Department of Electrical and Computer Engineering
Democritus University of Thrace

Algorithmic Analysis of User Behavior in Social Media

Giorgos Stamatelatos

A thesis submitted for the degree of
Doctor of Philosophy

Advisor: Pavlos S. Efraimidis

Xanthi, January 2022

Democritus University of Thrace
Department of Electrical and Computer Engineering
Building A, ECE, University Campus, 67100 Xanthi, Greece

To my parents.

Acknowledgements

The work research conducted in this thesis could not have been performed if not for the assistance, patience, and support of many individuals. Here, I would like to acknowledge their assistance and express my gratitude to them.

First and foremost, I offer my sincerest gratitude to my advisor, Pavlos Efraimidis, who has supported me throughout my thesis with his patience, knowledge and expertise while allowing me the room to work in my own way. His support on this research and advice was essential to its completion and the level of its quality. I would also like to thank the rest of my advisory committee members, Paul Spirakis and Vassilis Tsaousidis, as well as the members of my examining committee, Constantine Kotropoulos, Dimitris Fotakis, Avi Arampatzis and Eli Katsiri for their evaluation and comments on my work.

The research conducted within this thesis was partially supported by two research projects:

1. "Assessment of News Reliability in Social Networks of Influence" (Grant no. MIS 5006337) of the Operational Program "Human Resources Development, Education and Lifelong Learning" and is co-financed by the European Union (European Social Fund) and Greek National funds.

2. "Tools for the promotion of tourism experience" (project code: T1EDK-02474, grant no. MIS 5030446) which has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE.

I also wish to thank the members of these research projects, George Drosatos, Sotirios Gyftopoulos, Helen Briola, Vasileios Perifanis, for their collaboration in the joint works from these projects. I am finally willing to acknowledge the support of Chrysanthos Tassis and Costas Eleftheriou from the Department of Social Administration and Political Science, Democritus University of Thrace, in undertaking the survey in Section 3.

Abstract

Social networks are ubiquitous in our lives, the communities and the societies we participate in. Recently, the emergence of online social networks has triggered a major milestone in our interaction with our social circles. Online social networks have dramatically changed our social interactions and the way we seek information in the digital, interconnected world. Online users can interact with their family, their friends, or other users almost instantaneously, share their opinions and beliefs, and comment or react to other users' content. For this reason, online social networks play an increasingly significant role noawdays in both research and commercial applications. Naturally, most leading commercial entities are providing social connectivity services and compete for the users' online time, while researchers study a wide variety of components existing in social networks and online social networks, such as the behavior of users or utilizing the knowledge present in these massive databases to solve real problems.

In the context of structural analysis, social networks can be naturally thought as a collection of nodes that represent the entities that participate in them and a collection of links among them that represent their relationships. As a result, the activities performed by online users can also be thought as structural in more than one ways. One natural example would be the decisions of users to connect to other users based on real life social relationships or based on the content they choose to expose themselves to. A significant amount of research focuses on this structural data exclusively, with the most common being the direct friendship network among individuals, as a means of understanding how opinions, beliefs or other content propagates through the network. In addition, this knowledge and understanding can then be utilized in an attempt to solve real world problems that often occur within the context of online social networks. These concepts refer to social network analysis, which is the study of social networks by using graph theory as the underlying mechanism. The aforementioned tasks and problems, however, are sometimes challenging to address as they are based on studying the complex nature of human behavior in the online world.

In the work presented in this thesis we use both established and novel structural methodology social network analysis in order to recognize the online behavior of users and utilize the observations in order to attempt to solve problems that commonly arise in online social networks. Our assumption, that is confirmed –to an extent– through

our analysis, is that users will behave consistently with their interests, beliefs or tastes. This phenomenon gives rise to patterns and structure to real social networks as users have the tendency to create connections with other users or other entities based on criteria of interest, taste or exposure. Our methods are exclusively structural, i.e. they rely on the links formed in an explicit or implicit network between the relevant entities, which may not be necessarily formed among real users. In summary, our analysis is twofold: (1) structural analysis and (2) based on the expected behavior of users.

We focus our applications in problems that commonly arise in online social networks and are of particular interest to the context of study in this thesis: Greece. The first application being discussed is political affinity and, more specifically, applied to the multiparty, complex, political scene of Greece, as imprinted in social media. In this scenario, we apply methods in order to uncover the political affinity of important nodes of interest: the members of the Greek parliament (MPs) and the most popular news media active on social media. This topic is relevant as it is related to fundamental issues in online social networks, such as the reliability or bias in news stories. Another application discussed in this thesis is recommender systems in another area of important in Greece, which is tourism. We attempt to capitalize on the assumption that users will group related points of interest (POIs) together in order to infer the similarities among the POIs and power a recommendation system based on these similarities. Finally, we also study the preferential attachment mechanism, a fundamental mechanism that can explain certain properties of real social networks with respect to the decision of nodes to connect to other nodes. Here, the preferential attachment mechanism is studied in relation to the random sampling problem and a strictly correct and efficient implementation of the mechanism is developed as a growing random graph generator.

We have managed to utilize structural data from online social networks in order to extract useful knowledge using social network analysis methods. The knowledge is originating from the simple observation that the behavior of online users is consistent with their interests and beliefs, a phenomenon that creates structural patterns in the network. Our multi-perspective evaluations throughout this thesis supports this assumption. In particular, for the application of political affinity extraction, we managed to classify the vertices to their known political parties with surprising accuracy and arranged them into the left to right political axis using the minimum linear arrangement problem, whose application is novel in the context of social network analysis. For the case of tourism and the recommender system, we showed that there exists significant useful information in user defined point of interest lists. Our recommendation system achieved high effectiveness with respect to multiple evaluation scenarios and firmly surpassed the popularity baselines that are often considered difficult to exceed. The datasets and supplementary material in these works, such as the surveys, are new. We finally developed an algorithm that imprints the preferential attachment mechanism accurately and efficiently into the Barabási–Albert model and proved both its correctness and asymptotic performance.

Περίληψη

Τα μέσα κοινωνικής δικτύωσης κατέχουν κεντρικό ρόλο στη ζωή του σύγχρονου ανθρώπου και συντελούν σε μεγάλο βαθμό στη διαμόρφωση των ατόμων και των ομάδων. Βασικό αντικείμενο έρευνας και μελέτης στον τομέα των κοινωνικών δικτύων αποτελούν οι διεπαφές και διαπροσωπικές σχέσεις μεταξύ των εμπλεκόμενων οντοτήτων (ατόμων, ομάδων ή υποομάδων, οργανισμών ή άλλων μονάδων) σε διάφορα πεδία εφαρμογής (την πολιτική, την οικονομία, τη βιολογία, την κατανομή του πλούτου, τις συναλλαγές, την ψυχολογία, την επιστήμη των υπολογιστών κ.ά.) καθώς και η μελέτη της αμφίδρομης ροής της πληροφορίας ανάμεσα στις οντότητες αυτές. Η ανάλυση κοινωνικών δικτύων (social network analysis) είναι η μελέτη των κοινωνικών δικτύων που πραγματοποιείται μέσα από το γνωστικό αντικείμενο της θεωρίας γράφων (graph theory) καθώς και της θεωρίας δικτύων (network theory), που αποτελούν θεμελιώδη γνωστικά πεδία της επιστήμης δικτύων (network science). Η προσέγγιση αυτή αποσκοπεί στην πληρέστερη κατανόηση αλλά και τη βαθύτερη ερμηνεία των ποικίλων δεσμών, των αλληλεπιδράσεων ή των αμοιβαίων επικοινωνιακών σχέσεων μεταξύ ατόμων ή ομάδων, μια σειρά δηλαδή παραγόντων που με τη σειρά τους διαμορφώνουν και ρυθμίζουν (πέρα και έξω των γενετικών ή περιβαλλοντικών επιρροών) την ατομική συμπεριφορά ή την συλλογική/ομαδική συνείδηση και έκφραση.

Η συστηματική μελέτη των κοινωνικών δικτύων που καθιέρωσε την σύγχρονη ανάλυση κοινωνικών δικτύων εμφανίστηκε στα τέλη της δεκαετίας του 1990. Πληθώρα επιστημονικών μελετών αυτή την περίοδο συνέπεσε με την ραγδαία ανάπτυξη του Διαδικτύου καθώς και των υπηρεσιών άμεσης επικοινωνίας που εμφανίστηκαν σε αυτό. Οι συσχετίσεις ανάμεσα στις διασυνδεδεμένες οντότητες του αναδυόμενου ψηφιακού κόσμου ανέδειξαν την ανάπτυξη μεθόδων των δομικών ιδιοτήτων τέτοιων δικτύων αλλά και την κατανόηση των υποκείμενων μηχανισμών που διέπουν τη λειτουργία τους. Πολλά από τα θεμελιώδη είδη δικτύων που παρατηρούνται στις πραγματικές κοινωνίες των ανθρώπων ή ακόμη και στη φύση, όπως τα δίκτυα scale-free (Barabási & Albert, 1999) ή τα δίκτυα small world (Watts & Strogatz, 1998), μελετήθηκαν τη συγκεκριμένη περίοδο.

Το πιο πρόσφατο ορόσημο των κοινωνικών δικτύων και της ανάλυσής τους σημειώθηκε με την μετάβαση του πεδίου στα διαδικτυακά κοινωνικά δίκτυα (online social networks). Η εμφάνιση των διαδικτυακών κοινωνικών δικτύων δεν άλλαξε μόνο την ποσότητα και ποιότητα της διαθέσιμης πληροφορίας αλλά και επηρέασε σε μεγάλο βαθμό την διαδραστικότητα των χρηστών με αυτές τις πλατφόρμες. Δισεκατομμύρια χρήστες καθημερινά ενασχολούνται

με το περιεχόμενο διαφόρων υπηρεσιών κοινωνικής δικτύωσης, παράγουν χρήσιμη γνώση ενεργώντας με βάση τις προτιμήσεις τους, και αναζητώντας νέες πληροφορίες σχετικές με την παρουσία τους στην πλατφόρμα. Η διείσδυση τέτοιων κοινωνικών δικτύων στη ζωή μας είναι τόσο έντονη που η έννοια του κοινωνικού δικτύου είναι συχνά συνυφασμένη με τις διασυνδεδεμένες πλατφόρμες κοινωνικής δικτύωσης. Καθώς τα κοινωνικά δίκτυα έχουν δισεκατομμύρια ενεργούς χρήστες, τόσο οι επιπτώσεις αλλά και οι προοπτικές της χρήσης μιας τόσο πλούσιας δεξαμενής πληροφοριών είναι πολύπλευρες και εκτεταμένες.

Στην έρευνα που έχει διαξαχθεί στα πλαίσια αυτής της διατριβής, προσανατολιζόμαστε στην εξόρυξη χρήσιμης πληροφορίας και γνώσης από διαδικτυακά μέσα κοινωνικής δικτύωσης και την εκμεταλλευόμαστε για να λύσουμε προβλήματα που συχνά συναντώνται όταν οι χρήστες αναζητούν πληροφορίες σε αυτά. Η ανάλυσή μας περιβάλλεται από δύο βασικές ιδιότητες:

1. Βασίζεται σε μεθόδους που είναι καθαρά δομικές, δηλαδή εφαρμόζονται αποκλειστικά στην αναπαράσταση του δικτύου ως ένα σύνολο από κορυφές και ακμές.

2. Βασίζεται στην υπόθεση ότι οι διαδικτυακοί χρήστες παίρνουν αποφάσεις και ενεργούν με βάση την προσωπικότητά τους (τα ενδιαφέροντα και τις προτιμήσεις τους).

Συγκεκριμένα, εξετάζουμε εφαρμογές κοινωνικών φαινομένων που αφορούν την πολιτική και τον τουρισμό ως περιπτώσεις χρήσης. Επιπρόσθετα, επιχειρούμε να παρουσιάσουμε ένα γενικό θεωρητικό μοντέλο περιγραφής της δημιουργίας δικτύων που μπορεί εν δυνάμει να εξηγήσει ορισμένες από τις ιδιότητες των κοινωνικών δικτύων που αφορούν στην τάση των χρηστών να προσεγγίζουν συγκεκριμένους άλλους χρήστες στα δίκτυα αυτά.

Κατευθύνσεις και Κίνητρα

Η εμφάνιση των διαδικτυακών κοινωνικών μέσων άλλαξε δραματικά τον τρόπο με τον οποίο οι άνθρωποι επικοινωνούν και αλληλεπιδρούν. Τα κοινωνικά μέσα καθιέρωσαν ένα περιβάλλον στο οποίο η πρόσβαση και επαφή των χρηστών με την πλατφόρμα καθίσταται τετριμμένη με τη χρήση μοντέρνων διασυνδεδεμένων συσκευών. Επιπρόσθετα, οι πλατφόρμες κοινωνικής δικτύωσης εξελίσσονται διαρκώς με προσθήκες λειτουργιών, όπως μοίρασμα φωτογραφιών, τοποθέτηση ετικετών (tagging), πρόταση συστάσεων (recommendation) και άλλα. Στα κοινωνικά μέσα, οι χρήστες που συμμετέχουν παράγουν καθημερινά μεγάλη ποσότητα πληροφορίας. Η παραγωγή αυτή είναι συχνά διάφανη στους χρήστες οι οποίοι, μέσα από τις ενέργειές τους, συνεισφέρουν σε μια τεράστια δεξαμενή πληροφοριών που είναι συσχετισμένη με τις προτιμήσεις και τα ενδιαφέροντά τους. Ο ισχυρισμός μας είναι ότι η δομική συνιστώσα που εμπεριέχεται σε αυτή την πληροφορία πηγάζει από τα κίνητρα που δίνονται στους χρήστες για τη συμμετοχή τους στο κοινωνικό μέσο και μπορούν να χρησιμοποιηθούν για τη μελέτη μιας εφαρμογής στο κοινωνικό αυτό δίκτυο. Στην έρευνα αυτή, εξετάζουμε δύο σημαντικές περιπτώσεις: γνώση πολιτικού περιεχομένου και συναφής γνώση για τουριστικές εφαρμογές.

Πολιτικό περιεχόμενο Τα ζητήματα πολιτικής σημασίας αποτελούν αντικείμενο διαρκούς απασχόλησης ενώ την σύγχρονη εποχή εμφανίζονται συχνά στα κοινωνικά μέσα. Καθώς τα κοινωνικά μέσα συνεχίζουν να κατέχουν κεντρικό ρόλο στις πολιτικές καμπάνιες, φαινόμενα ψευδών ειδήσεων (fake news) ή πολιτικής προκατάληψης (bias) άρχισαν να κάνουν την εμφάνισή τους και να διαδίδονται μέσω κοινωνικών κύκλων στα κοινωνικά μέσα. Η μελέτη τέτοιων φαινομένων κρίνεται ιδιαίτερα σημαντική λόγω του αντίκτυπου που έχουν στην κοινωνική πόλωση καθώς και στους θεμελιώδεις μηχανισμούς της δημοκρατίας. Κρίνεται συνεπώς σημαντική η αυτόματη εξαγωγή πολιτικής πληροφορίας σχετικής με συγκεκριμένους κόμβους ενδιαφέροντος υψηλής κεντρικότητας σε ένα δίκτυο, όπως για παράδειγμα ο πολιτικός τους προσανατολισμός ή η τάση τους να παράγουν ψευδές ή προκατειλημένο περιεχόμενο. Χρησιμοποιώντας όμως τη συμπεριφορά και τις ενέργειες των ιδίων για τον αυτόματο εντοπισμό αυτής της πληροφορίας ίσως είναι πιο απαιτητικό να επιτευχθεί καθώς οι χρήστες έχουν τη δυνατότητα να προσαρμόσουν τη συμπεριφορά τους για να αποφύγουν μία τέτοια ενέργεια. Ο ισχυρισμός μας είναι ότι συχνά είναι πιο αξιόπιστος ο εντοπισμός του πολιτικού προσανατολισμού ενός χρήστη μέσω των συνδέσεών του με άλλους χρήστες. Συγκεκριμένα, χρησιμοποιούμε τους συνδέσμους που σχηματίζονται από άλλους χρήστες προς τους χρήστες ενδιαφέροντος, οι οποίοι μπορεί να μην είναι αμοιβαίοι. Με αυτόν τον τρόπο, ο πολιτικός προσανατολισμός των χρηστών ενδιαφέροντος εντοπίζεται με βάση τη συλλογική τους εικόνα στο δικτύου, και όχι μόνο με βάση του ατομικού τους προφίλ.

Προσωποποιημένες συστάσεις Μια ακόμη ενδιαφέρουσα εφαρμογή ανάλυσης κοινωνικών δικτύων που είναι μέρος αυτής της διατριβής είναι τα συστήματα προσωποποιημένων συστάσεων (recommender systems) στα κοινωνικά μέσα που ειδικεύονται στον τουρισμό: τα τουριστικά κοινωνικά μέσα. Σε αυτή την περίπτωση, η υπόθεσή μας είναι πως η συμπεριφορά των χρηστών του κοινωνικού μέσου θα είναι συνεπείς με τις προτιμήσεις και τα ενδιαφέροντά τους. Ένας φυσικός τρόπος για υπόδειξη προσωποποιημένων συστάσεων είναι η εκμετάλλευση της εγγενούς πληροφορίας στα προφίλ των χρηστών που συμμετέχουν ενεργά στο κοινωνικό μέσο και προέρχονται από τις ενέργειες των ίδιων. Οι μέθοδοί μας χρησιμοποιούν δημόσιες συλλογές αντικειμένων που βασίζονται στο ότι οι χρήστες συχνά επιθυμούν να επισκέπτονται περισσότερους προορισμούς από ότι πραγματικά επισκέπτονται. Βασιζόμενοι σε αυτή την ιδιότητα, επιχειρούμε να εντοπίσουμε τις ομοιότητες ανάμεσα στις τοποθεσίες, έτσι όπως αυτή είναι αποτυπωμένη από τους χρήστες του κοινωνικού μέσου.

Παραγωγή γράφων scale-free Το τελευταίο μέρος της διατριβής είναι ένα θεωρητικό μοντέλο δημιουργίας γράφων που παρουσιάζει ομοιότητες με διαδικασίες που εμφανίζονται σε πραγματικά κοινωνικά, ή μη-κοινωνικά, δίκτυα, όπως βιολογικά δίκτυα ή δίκτυα υπολογιστών. Πιο συγκεκριμένα, η ιδιότητα αυτή αναφέρεται στα scale-free δίκτυα, δηλαδή τα δίκτυα των οποίων η κατανομή των βαθμών ακολουθεί κατανομή νόμου δύναμης (power law). Υπάρχουσες μελέτες στην βιβλιογραφία δείχνουν ότι η τωρινή κατάσταση στον χώρο δεν περιγράφει κάποιο μοντέλο που να είναι ακριβές σε σχέση με το πιθανοκρατικό μοντέλο που χρησιμοποιείται αλλά καλές προσεγγίσεις αυτού. Για τον λόγο αυτό, παρουσιάζουμε

ένα μοντέλο που υλοποιεί επακριβώς τον εγγενή μηχανισμό της δημιουργίας του δικτύου χωρίς να περιορίζουμε τον χρόνο εκτέλεσης της διαδικασίας.

Μεθοδολογία

Η ανάλυση που γίνεται στα πλαίσια της διατριβής αυτής βασίζεται σε μεθόδους που είναι αποκλειστικά δομικές, δηλαδή χρησιμοποιούν ως πρωτογενή πληροφορία μόνο δομικά χαρακτηριστικά του κοινωνικού δικτύου. Πιο συγκεκριμένα, η δομή ενός κοινωνικού δικτύου ορίζεται ως ένα σύνολο κόμβων που χαρακτηρίζουν τις οντότητες που συμμετέχουν στο κοινωνικό δίκτυο, και ένα σύνολο από ακμές που αντιπροσωπεύουν τις σχέσεις ανάμεσα στους κόμβους. Η ανάλυσή μας βασίζεται αποκλειστικά σε αυτό το είδος πληροφορίας και δεν λαμβάνεται υπόψη άλλος τύπος γνώσης σχετικός με τους κόμβους του δικτύου ή τις σχέσεις ανάμεσά τους. Επιπρόσθετα, οι μέθοδοι που χρησιμοποιούμε συνήθως μεταχειρίζονται τις ακμές ως δυαδικές σχέσεις ύπαρξης ή απουσίας σχέσης ενώ οι κόμβοι χαρακτηρίζονται με ένα αναγνωριστικό (ID) χωρίς φυσική ερμηνεία ή αξία. Υπό αυτό το πρίσμα, η ανάλυση κοινωνικών δικτύων θεωρείται το κύριο αντικείμενο αυτής της διατριβής σε ό,τι αφορά τη μεθοδολογία. Στην έρευνά μας χρησιμοποιούμε τόσο καθιερωμένες μεθόδους ανάλυσης του χώρου για την εξόρυξη της απαιτούμενης πληροφορίας από το δίκτυο όσο και καινοτόμες μεθόδους που δεν έχουν εφαρμοστεί στα κοινωνικά μέσα, και μελετάμε τη φυσική ερμηνεία της εφαρμογής τους.

Η συχνότερη, καθιερωμένη στη βιβλιογραφία, μέθοδος που χρησιμοποιείται στην ανάλυσή μας είναι η μέθοδος graph projection, που αποτελεί έναν μετασχηματισμό του γράφου που μπορεί να συλλάβει ως ένα βαθμό τις σχέσεις ανάμεσα σε ζευγάρια κόμβων του γράφου, οι οποίες δεν μπορούν να αποτυπωθούν με τη χρήση των άμεσων σχέσεων στον αρχικό γράφο. Οι μέθοδοι graph projections απλοποιούν ένα δίκτυο, χωρίς να περιορίζουν σημαντικά την πληροφορία που υπάρχει σε αυτό, και κατευθύνουν την ανάλυση σε μια συγκεκριμένη επιλεγμένη ομάδα κόμβων. Άλλες μέθοδοι ανάλυσης κοινωνικών δικτύων αποτελούν οι τυχαία περίπατοι (random walks), η συσταδοποίηση κόμβων (community detection), ή άλλες μέθοδοι σημασιολογικής περιγραφής των οντοτήτων που συμμετέχουν στο δίκτυο. Σε αρκετές περιπτώσεις, η έρευνά μας περιλαμβάνει συγκριτική παρουσίαση αυτών των μεθόδων και των παραμέτρων τους. Τέλος, χρησιμοποιούμε μεθόδους αλγοριθμικής ανάλυσης δικτύων για καινοτόμα εφαρμογή στα κοινωνικά μέσα και δείχνουμε την καταλληλότητά τους ως προς την ανάδειξη της φυσικής ερμηνείας της πρωτογενούς πληροφορίας που υπάρχει στο δίκτυο.

Η μεθοδολογία μας εφαρμόζεται σε πρωτογενή δεδομένα από τα κοινωνικά δίκτυα του Twitter και Foursquare. Τα δεδομένα που υπάρχουν και στις δύο αυτές υπηρεσίες ταιριάζουν απόλυτα με τα κίνητρα και τους στόχους μας καθώς και με την περιγραφή των προβλημάτων που επιχειρούμε να δώσουμε λύση σε αυτή την έρευνα. Το Twitter συχνά αντιμετωπίζεται ως ένα κοινωνικό μέσο με έντονη την παρουσία της πολιτικής συνιστώσας (Parmelee, 2014) και αποτελείται από ένα πλούσιο σύνολο ενεργών χρηστών, όπως

πολιτικοί, αντιπρόσωποι κομμάτων, υποψήφιοι, ή ακόμη και μέσα ενημέρωσης. Χρησιμοποιώντας ανάλυση κοινωνικών δικτύων, δείχνουμε την καταλληλότητα του Twitter για την ανάλυση φαινομένων πολιτικής προκατάληψης. Από την άλλη μεριά, το Foursquare είναι μια πλατφόρμα τουριστικού βοηθού και δεδομένων τοποθεσίας με περίπου 50 εκατομμύρια χρήστες. Οι χρήστες αλληλεπιδρούν με άλλους χρήστες ή με σημεία ενδιαφέροντος, χαρακτηριστικό του Foursquare που το κάνει ιδανικό για τη μελέτη προσωποποιημένων συστάσεων σημείων ενδιαφέροντος. Και τα δύο κοινωνικά μέσα αποτελούνται από ένα μεγάλο αριθμό χρηστών που έχουν κίνητρα να συμπεριφέρονται με συγκεκριμένο τρόπο, να παίρνουν αποφάσεις και να ενεργούν με βάση τις προτιμήσεις, τα ενδιαφέροντα και τις πεποιθήσεις τους. Στην έρευνα αυτή, επιχειρούμε να εκμεταλλευτούμε αυτό το φαινόμενο, συναθροίζοντας τις συμπεριφορές πολλαπλών χρηστών για την παραγωγή γνώσης επί τους αντικειμένου της μελέτης. Η ιδιότητα αυτή είναι βασισμένη στο γεγονός ότι οι χρήστες συχνά επιλέγουν να εκθέτουν τον εαυτό τους σε πληροφορίες που είναι σχετικές με τις προτιμήσεις τους ή ενδυναμώνουν τις απόψεις τους.

Τέλος, παρουσιάζουμε μία αλγοριθμική προσέγγιση δημιουργίας κοινωνικού γράφου που επιχειρεί να εξηγήσει μία από τις θεμελιώδεις ιδιότητες πάνω στην οποία βασίζεται η δημιουργία σχέσεων σε κοινωνικά ή ακόμη και μη-κοινωνικά δίκτυα στον πραγματικό κόσμο. Η ιδιότητα αυτή αναφέρεται ως ο μηχανισμός του preferential attachment, σύμφωνα με τον οποίο οι κόμβοι που εισέρχονται στο δίκτυο συνήθως συνδέονται με παλαιότερους κόμβους που έχουν ήδη υψηλή συνδεσιμότητα. Οι αλγόριθμοι που προτείνουμε είναι αναλυτικές μέθοδοι αποδοτικής εκτέλεσης που συμπεριλαμβάνουν αυτή τη διάσταση της αλληλεπίδρασης μεταξύ των χρηστών.

Σύνοψη

Κεφάλαιο 2: Υπόβαθρο Παρουσιάζονται σύντομες περιγραφές των αντικειμένων που εξετάζονται στη διατριβή. Εξετάζεται η αναπαράσταση των δικτύων ως γράφων μαζί με τις μεθόδους που συχνά χρησιμοποιούνται για τη μελέτη μιας τέτοιας δομής. Δίνεται η περιγραφή των δικτύων scale-free και του μηχανισμού preferential attachment, που είναι μία από τις πιο δημοφιλείς εξηγήσεις σχετικά με το πώς σχηματίζονται τα δίκτυα scale-free. Επιπρόσθετα, περιγράφονται μέθοδοι που χρησιμοποιούνται συχνά στη συνέχεια της διατριβής, οι οποίες χρησιμοποιούν και απεικονίζουν την πληροφορία που εξάγεται από τα κοινωνικά μέσα. Τέλος, αναφέρουμε το πρόβλημα της τυχαίας δειγματοληψίας (random sampling) καθώς είναι θεμελιώδες σε ορισμένες διεργασίες που διέπουν τη συμπεριφορά των χρηστών στα κοινωνικά δίκτυα.

Κεφάλαιο 3: Πολιτικός Προσανατολισμός στο Twitter Στο κεφάλαιο αυτό, δείχνουμε ότι τα δομικά χαρακτηριστικά του κοινωνικού δικτύου του Twitter μπορούν να αποκαλύψουν πολύτιμες πληροφορίες σχετικές με τον πολιτικό προσανατολισμό των ενεργών οντοτήτων. Πιο συγκεκριμένα, δείχνουμε ότι οι ακόλουθοι (followers) στο Twitter μπορούν να χρησιμοποιηθούν για να προβλέψουμε τον πολιτικό προσανατολισμό άλλων χρηστών

που επιλέγουν να ακολουθήσουν. Χρησιμοποιούμε ένα σύνολο από αποκλειστικά δομικές αλγοριθμικές προσεγγίσεις για να αποκαλύψουμε πολλαπλές συνιστώσες του πολιτικού προφίλ των χρηστών. Οι μέθοδοί μας εφαρμόζονται σε ένα σύνολο δεδομένων από την Ελληνική πολιτική σκηνή και τα αποτελέσματα επιβεβαιώνουν τους ισχυρισμούς μας και την ορθότητα της προσέγγισης.

Κεφάλαιο 4: Προσωποποιημένες Συστάσεις μέσω Λιστών Foursquare Στο κεφάλαιο αυτό, εξετάζουμε την πληροφορία που εμπεριέχεται σε μια ιδιαίτερη πηγή δομικών δεδομένων, τις λίστες ή συλλογές αντικειμένων, και εκτιμούμε τη δυνατότητα να εφαρμοστούν σε συστήματα προσωποποιημένων συστάσεων. Η υπόθεσή μας είναι ότι η πληροφορία που εμπεριέχεται στις λίστες μπορεί να χρησιμοποιηθεί για να εκτιμηθεί η ομοιότητα ανάμεσα στα αντικείμενα, ενώ οι ομοιότητες αυτές μπορούν με τη σειρά τους να οδηγήσουν ένα σύστημα προσωποποιημένων συστάσεων. Η αρχή λειτουργίας του συστήματος είναι το γεγονός ότι οι λίστες είναι συλλογές που δημιουργούνται από τους ίδιους τους χρήστες του δικτύου και, ως εκ τούτου, βασίζονται στην κρίση και επιλογές των χρηστών. Ως αποτέλεσμα, οι λίστες μπορούν να θεωρηθούν ως συλλογές συσχετισμένων αντικειμένων. Οι μέθοδοί μας επιχειρούν να εξάγουν αυτές τις συσχετίσεις και να εκφράσουν την έννοια της ομοιότητας μέσω γραφοθεωρητικών, συνολοθεωρητικών και στατιστικών μέτρων. Η προσέγγισή μας εφαρμόζεται σε ένα σύνολο δεδομένων από το Foursquare που περιλαμβάνει δύο δημοφιλείς τουριστικούς προορισμούς στη Βόρεια Ελλάδα ενώ τα αποτελέσματα επιβεβαιώνουν την ύπαρξη πλούσιας πληροφορίας ομοιότητας στις λίστες καθώς και την αποδοτικότητα της προσέγγισης σαν σύστημα συστάσεων.

Κεφάλαιο 5: Αποδοτική Παραγωγή Γράφων Scale-Free Το κεφάλαιο αυτό παρουσιάζει την ανάπτυξη μιας νέας ομάδας αλγορίθμων που υλοποιεί επακριβώς τον μηχανισμό του preferential attachment, την πιο διαδεδομένη μέθοδο κατασκευής γράφων scale-free που εμφανίζονται στη φύση και στις ανθρώπινες κοινωνίες. Σε αντίθεση με τωρινά, προσεγγιστικά σχήματα του μηχανισμού αυτού, οι μέθοδοί μας είναι ακριβείς ως προς την αναλογικότητα των πιθανοτήτων επιλογής κόμβων με τους βαθμούς των κόμβων αυτών, και εκτελούνται σε γραμμικό χρόνο ως προς το πλήθος των κόμβων του γράφου που προκύπτει. Η προσέγγισή μας βασίζεται σε ένα συνδυασμό μεθόδων random sampling, των οποίων η εφαρμογή στο πρόβλημα του preferential attachment είναι καινοτόμα.

Κεφάλαιο 6: Εργαλεία Λογισμικού Στο κεφάλαιο αυτό παρουσιάζονται δύο εργαλεία λογισμικού που αναπτύχθηκαν κατά τη διάρκεια εκπόνησης της έρευνας της διατριβής, οι βιβλιοθήκες `random-sampling` και `social-influence`. Συγκεκριμένα, η βιβλιοθήκη `random-sampling` είναι μια συλλογή από υλοποιήσεις αλγορίθμων δειγματοληψίας reservoir, τόσο με βάρη όσο και χωρίς βάρη, όπου η απαίτηση μνήμης όλων των υλοποιήσεων είναι γραμμική ως προς το μέγεθος του δείγματος. Η βιβλιοθήκη `social-influence` είναι μια ευρύτερη συλλογή από εργαλεία και αλγορίθμους που αφορούν την υλοποίηση δομών δεδομένων γράφων, εργαλεία για ανάλυση κοινωνικής συμπεριφοράς, κοινωνικά μοντέλα επιρροής και άλλα.

Κεφάλαιο 7: Συμπεράσματα Τα συνολικά συμπεράσματα αυτής της δουλειάς παρουσιάζονται σε αυτό το κεφάλαιο. Η δουλειά της διατριβής αυτής συνοψίζεται και εξετάζονται πιθανές μελλοντικές κατευθύνσεις σχετικές με το αντικείμενο.

Contents

List of Tables

List of Figures

List of Abbreviations

ADA Americans for Democratic Action. 36

API Application Programming Interface. xix, 34, 38, 68, 74, 84, 91, 127–129, 132, 133, 136, 137

BA Barabási–Albert. viii, 105–109, 115, 118, 119, 124, 125, 128, 136, 142

EM Expectation Maximization. 56

GCD Gratest Common Divisor. 24

GPS Global Positioning System. 66

HITS Hyperlink-Induced Topic Search. 137

LA Linear Arrangement. xxiii, 43, 44, 50–52

LBSN Location Based Social Network. 3, 25, 65, 66, 68–71, 91, 93, 102, 140

LCM Least Common Multiple. 137

MI Mutual Information. 69, 79, 88, 90, 100

MICE Multivariate Imputation by Chained Equations. 56

MinLA Minimum Linear Arrangement. viii, xxi, xxiv, 20, 21, 23, 31, 34, 35, 42–45, 47, 50–52, 54, 56–58, 62, 63, 131, 132, 137, 140, 142

MP Member of Parliament. viii, xviii, xxi, xxiii, 32, 34, 36–38, 45–55, 57–63, 140, 142

NMI Normalized Mutual Information. 48

NOI Node of Interest. xxi, xxiii, 31–33, 36–45, 47, 53–57, 59–63, 140

OSN Online Social Network. vii, viii, 1–3, 7, 10, 12, 13, 17, 25, 31, 63, 66, 139

POI Point of Interest. viii, xviii, xxii, xxiv, xxv, 25, 65–86, 88–94, 96–100, 102, 103, 139–141

PTAS Polynomial Time Approximation Scheme. 23

SMC Simple Matching Coefficient. 48, 138

SNA Social Network Analysis. vii, viii, 1, 3–5, 9, 11, 18, 34, 63, 132, 133, 139, 142

tf–idf term frequency-–inverse document frequency. 95

UGC User Generated Content. 65, 66, 69, 102, 142

VLSI Very Large Scale Integration. 23, 43

WRS Weighted Random Sampling. xvii, 27–29, 105–107, 128, 129

WS Watts–Strogatz. 136

Chapter 1

Introduction

Social networks are ubiquitous. They are central to individuals' lives and they shape the evolution of communities. Social networks are a relational construct that is used to study the relationships between social entities in various fields, such as finance, politics, economy, biology, health, communication, phychology, computer science and others. They are made up of entities, which could be individuals, organizations or other units, and the relationships among them, where the relationships represent some amount of information flow between the relevant entities. Social network analysis (SNA) is the study of social networks through the use of graph and network theory and is a fundamental field of network science for understanding and investigating social structures.

The systematic study of social networks that established the modern version of social network analysis appeared in the late 1990s. Various such studies have coincided with the Internet's growth as well as the emergence of online communication services within it. The relationships among the connected entities in the emerging digital world have given rise to the development of methods of structural analysis of these networks as well as the understanding of the underlying processes that create them. Certain important and fundamental types of networks that can be observed in human societies and nature, such as scale-free (Barabási & Albert, 1999) or small world (Watts & Strogatz, 1998) networks, were identified during this period.

More recently, another milestone of social networks and social network analysis has been marked by the development of online social networks (OSNs). The emergence of online social networks (interchangeably used with *social media*) has not only dramatically changed the quantity and quality of available information but also expanded the role of these services in our lives. Every day, billions of active social media users engage in content in various online platforms, generate knowledge following their actions and online presence, and seek novel information relevant to them. The ubiquity of social media is so prevalent that nowadays, the terms social network and online social network are interwoven and often used interchangeably. As OSNs have billions of active users

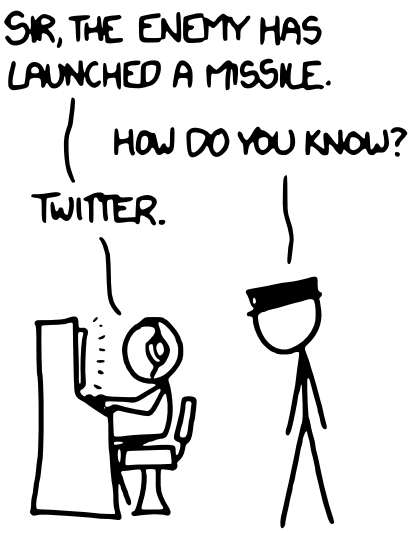


Figure 1.1: - Sir, the enemy has launched a missile. - How do you know? - Twitter. (Munroe, 2014, xkcd.com).

online, both the implications and the potential of utilizing such rich database of information are numerous and far reaching. This potential has been comically illustrated by Randall Munroe in the "Alien Astronomers" (Munroe, 2014, xkcd.com). A copy of the illustration is in Figure 1.1.

In the research conducted in this thesis, we aim to extract useful knowledge and insights from online social networks and use them to solve common problems that arise when users seek information within them. Our analysis is surrounded by two core properties:

1. It is based on methods that are purely structural, i.e. rely exclusively on the network link structure.
2. It is based on the assumption that online users tend to make decisions and take actions consistent with their beliefs or tastes.

In particular, we examine applications of social phenomena in politics and tourism as use cases. We also attempt to provide a general theoretical model of growing network formation that can explain certain properties of social networks regarding the tendency of users to connect to other users in networks.

1.1 Directions and Motivation

The emergence of online social networks has revolutionized the way people socialize, interact and communicate with each other. OSNs established an environment where the access and interaction of people with the platform is trivial and can be performed via any of the connected devices owned by users. Furthermore, these social networking platforms have evolved to include increasingly higher dimensions of connectivity, such as photo sharing, topic tagging, recommendations and others. In these online social

networks, the users are generating large volumes of information daily. Often, users might not even realize that, through their actions, they are contributing to a massive database of data that is correlated with the beliefs and tastes of individual users. We argue –and demonstrate– that the structural dimension of this information originates from the motivation given to the users to access the online platform and can be used to empower an application for the context of the social features studied in that network. In this research, we use two prominent use cases: political knowledge and contextual knowledge for tourism applications.

Political content Political issues are historically relevant and nowadays constantly arise in online social networks. As OSNs continue to gain central roles in political campaigns, phenomena of fake news or politically biased news have started to appear and propagate through online social circles. These phenomena are particularly relevant, especially through their impact in social polarization as well as the fundamental mechanisms of democracy. It is, therefore, important to be able to automatically extract political information about specific nodes of high importance in a network, for example their political standing or their tendency to create false or biased content. Relying, however, on their individual actions or profiles to determine this information might be more challenging to achieve as users might adapt their behavior in an attempt to avoid such detections. We argue that it is often more reliable to determine the political standing of individuals via their links to other users. In particular, we utilize the links that originated from other users and might not be reciprocal. In this way, we determine the political profile of individual users as imprinted by the whole network, not necessarily by the individuals' own actions.

Personalized recommendation Another interesting area of application for social network analysis that is part of this research is recommendation systems in online social media that are special types of OSNs in tourism, namely the Location Based Social Networks (LBSNs). In this use case, our assumption is also that the behavior of users will be consistent with their tastes. Leveraging the information that is embedded in the profiles of the users that participate in the network and is manifestated through their actions is a natural way to provide location recommendations. Our methods leverage crowd sourced collections of items that follow the premise that users often want to visit more places that they actually do. Based on this property, we are able to infer the pairwise similarities among the locations, as it is imprinted by the users of the network themselves.

Scale-free network formation The last part of the research in this thesis is a theoretical model of network formation that resembles processes that appear naturally in real social or non-social networks, such as biological networks or computer networks. In particular, this is referring to scale-free networks; networks whose degree distribution follows a power law. Literature reviews indicate that current algorithmic

models are not precise in terms of the probabilistic model employed but very good approximations of it. For this reason, we establish a model that accurately implements the underlying mechanism of network formation without comprimising the running time performance.

1.2 Methodology

The analysis performed within this thesis is based on methods that are purely structural, i.e. they rely on the network structure exclusively. In particular, the network structure is defined as a set of nodes that represent the actors of the social network, and edges that represent their relations. Our analysis relies on this structure alone and other non-structural information about the entities of the network or their relationships are not taken into consideration. Furthermore, our methods often treat the edges simple as binary relations among the nodes and the nodes themselves are modeled using semantically irrelevant information, for example a serial number or an ID. To that end, social network analysis (SNA) is considered the main topic in this thesis with respect to the methodology. While established SNA methods are utilized to derive the required knowledge from the networks, novel techniques that haven't been applied to social network analysis, but fit the description of individual problems, are used in order to extract knowledge related to the interpretation of physical processes.

A recurrent theme of established methodology applied in the research of this thesis are the graph projections, which are transformations of the network that can capture some amount of the pairwise relations among its entities in a higher dimension, beyond the direct relationships. Projections are important tools that simplify a network, making the study of networks heavily skewed towards a particular set of entities possible. Other methods of social network analysis include random walks, community detection or various other graph-theoretic appraches that can semantically describe the entities involved in the network. In several occasions, the research includes comprehensive comparison between these methods or different parameter settings. Finally, we utilize novel algorithmic approaches in the field of social network analysis and demonstrate their suitability and their ability to physically interpret the information within social networks.

Our methodology is applied on datasets from the Twitter and Foursquare social networks. Each of these portals can perfectly capture our motivation and goals and fit the descriptions of individual problems and questions posed in this research. Twitter has often been described as a medium of political advocacy and deliberation (Parmelee, 2014) with a rich set of politically active users, such as politicians, party representatives, candidates or news media. Using SNA, we demonstrate the suitability of the Twitter network for the analysis of political bias. On the other hand, Foursquare is a travel city guide and location data platform consisting of about 50 million connected users. Users

can interact with other users or with locations, making Foursquare an ideal medium to study location recommendations. Both social networks consist of a large user base that are motivated to behave in a certain way, make decisions, and take actions that are dictated by their tastes or their beliefs. In this research, we attempt to exploit this behavior and aggregate the behavior of multiple users to establish knowledge surrounding the topic of study. This property is based on the fact that users often choose to expose themselves to information that appears relevant to their interests or reinforcing to their beliefs.

Finally, we present an algorithmic approach of growing network formation that attempts to explain one of the fundamental properties of user connections in some social or non-social networks of the real world. This property is referring to the preferential attachment mechanism, via which new nodes in a network usually connect to other nodes that are already highly connected. Our proposed algorithms are analytical methods that can efficiently and accurately capture this dimension of user interaction.

1.3 Synopsis of Results

1. **Deriving the political affinity of twitter users from their followers (Stamatelatos et al., 2018)**.

   SNA methodology is applied on a Twitter dataset to establish that the decisions of Twitter users can portray the political orientation of members of the Greek parliament.

   - Core work of this PhD research.

2. **Revealing the political affinity of online entities through their Twitter followers (Stamatelatos et al., 2020)**.

   SNA methodology is extended on an enriched Twitter dataset to demonstrate the effectiveness on popular news media as well. News media are classified in terms of political bias.

   - Core work of this PhD research.

3. **A Twitter-based approach of news media impartiality in multipartite political scenes (Gyftopoulos et al., 2020)**.

   The notion of impartiality is being studied in the multipartite politice scene of Greece to determine the presence of political bias in popular news media. The news sources are ranked according to their political impartiality.

   - Joint work. Contribution of this PhD research: Collection and proprocessing of primitive data from the Twitter social network for the experimental setup.

Suggestions of projection measures and consulting on the interpretation of the results.

4. **Point-of-interest lists and their potential in recommendation systems (Stamatelatos et al., 2021)**.

   A recommendation system based on the novel use of item lists. Item lists are collections of items that are automatically created based on decisions of users in social networks and, therefore, contain knowledge about the similarities of the items.

   - Core work of this PhD research.

5. **Privacy leakages about political beliefs through analysis of Twitter followers (Briola et al., 2018)**.

   Privacy leakages about the political orientation of Twitter users are being examined based on the follower and followee connections with other users.

   - Joint work. Contribution of this PhD research: Collections of suitable data for studying the potential leakage of political information from the connections of a user in social networks. Observations related to the experimental results and their interpretation in relation to the actual data of the social network.

6. **Whole Sampling Generation of Scale-Free Graphs (Stamatelatos & Efraimidis, 2021b)**.

   Analytical presentation of a new scale-free network generator whose probability model is efficient and strict in terms of the proportionality of selection with the node degrees.

   - Core work of this PhD research.

7. **Datasets and software tools**.

   All primitive and preprocessed data that have been collected for the purposes of the work contained within this thesis have been made publicly available. This PhD research has resulted in two software libraries being developed. All software tools and frameworks that have been developed have been made publicly available in online code repositories.

1.4 Overview of the Thesis

Chapter 2: Background Brief descriptions of the topics involved in this thesis are presented. The representation of networks via graphs is examined along with the

methods that are commonly used to study this structure. The description of scale-free networks and the preferential attachment mechanism, one of the most populalar explanations of how scale-free networks form, is being given. We also describe the methods that commonly occur in the work presented in this thesis that utilize and visualize the information contained within social networks. Finally, we provide a brief description of the random sampling problem as it is fundamental to the understanding of certain processes that drive decisions made by users in networks.

Chapter 3: Political Affinity on Twitter In this chapter, we show that the structural features of the Twitter OSN can divulge valuable information about the political affinity of the participating nodes. More precisely, we show that Twitter followers can be used to predict the political affinity of prominent users of political importance they opt to follow. We utilize a series of purely structure-based algorithmic approaches in order to reveal diverse aspects of the users' political profile. Our methods are applied to a Greek political dataset and our results confirm the viability of our approach.

Chapter 4: Recommendation System via Foursquare Lists In this chapter, we investigate the information contained in unique structural data of OSNs, namely the *lists* or *collections* of items, and assess their potential in recommendation systems. Our hypothesis is that the information encoded in the lists can be utilized to estimate the similarities amongst items and, hence, these similarities can drive a personalized recommendation system. This is based on the fact that item lists are user generated content and, as such, are based on the networks users' judgement and decisions. As a result, they can be considered as collections of related items. Our method attempts to extract these relations and express the notion of similarity using graph theoretic, set theoretic and statistical measures. Our approach is applied on a Foursquare dataset of two popular destinations in northern Greece and the results confirm the existence of rich similarity information within the lists and the effectiveness of our approach as a recommendation system.

Chapter 5: Whole Sampling Generation of Scale-Free Graphs This chapter presents the development of a new class of algorithms that accurately implement the preferential attachment mechanism, the most commonly used method to create scale-free networks that appear in nature and human societies. Contrary to existing approximate preferential attachment schemes, our methods are exact in terms of the proportionality of the vertex selection probabilities to their degree and run in linear time with respect to the order of the generated graph. Our algorithms are based on a combination of random sampling methods, whose application is novel in the preferential attchment problem.

Chapter 6: Software Frameworks In this chapter, two frameworks that have been developed in the context of this thesis are presented, namely `random-sampling` and

`social-influence` libraries. In particular, `random-sampling` is a collection of implementations for reservoir sampling algorithms, both weighted and unweighted, where the memory consumed by each implementation is linear with respect to the size of the required sample. The `social-influence` library is a broader collection of tools and algorithms that range from the implementation of graph structures to tools for social behavior analysis or influence social models.

Chapter 7: Conclusions The overall conclusions of this work are presented in this chapter. The work presented in this thesis is summarized and possible future directions regarding the field are discussed.

Chapter 2

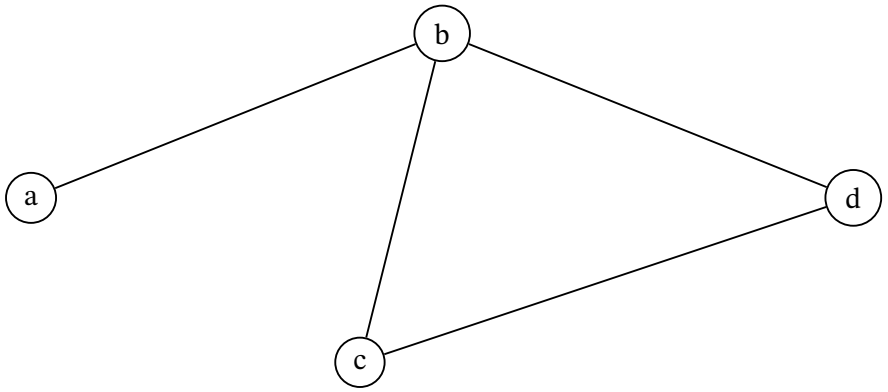
Background

This thesis is concerned with the study of social networks and the behavior of users in them. In the following sections of this chapter, brief descriptions of the fundamental topics relevant to the study of the thesis are presented. In particular, it is being initially examined how networks are represented using graphs along with an overview of the methods used to study social networks in the macroscopic level. A special category of networks that occur commonly in nature is described as the scale-free networks along with the mechanism of preferential attachment, which is one of the most popular explanations of how these networks form. Moreover, we describe the methods that commonly occur in the work presented in this thesis that utililize and visualize the information contained within social networks. Finally, we provide a brief description of the random sampling problem as it is fundamental to the understanding of certain processes that drive decisions made by users in networks.

2.1 Representing Social Networks

Social structure present in networks can naturally be represented as a graph. Graphs are flexible data structures that can capture many dimensions of the information contained within social networks. More formally, a graph $\mathcal{G}(N, E)$ is a set of nodes or vertices N and a set of edges or links E, where every edge corresponds to some kind of relationship between two vertices. For example, an edge $e_{xy} = (x, y)$ represents a relationship between the vertices x and y, which is typically represented with a numeric value (weighted graphs). The most common type of graph in social network analysis is the simple graph, where there may not be multiple edges between the same pair of nodes. Often, in these graphs, no loops are allowed either, i.e. no edges representing a relationship between the same vertex, i.e. (x, x).

The entities that take part in a social network comprise the set of nodes $N = \{1, 2, \dots, n\}$ participating in its graph representation. Often, nodes are the individuals participating



	a	**b**	**c**	**d**
a	0	1	0	0
b	1	0	1	1
c	0	1	0	1
d	0	1	1	0

Figure 2.1: A toy example of a graph representing a social network comprising 4 nodes. The visual illustration is shown on the left while the respective adjacency matrix is displayed on the right side.

in the social network (or the user accounts in social media) but might also represent other entities, such as countries, or even implicit data, for example categorical data. These entities are the actors of the social network and are represented as vertices in the graph structure.

The edges of the graph represent the social relationships between the vertices (the actors of the social network). Typically the type of relationship conveyed by edges are social relations or social interactions, such as friendships. This is often the case when we study online social networks, where the edges correspond to friendships in the network. Classic examples of these types of relationships can be found in the Renaissance Florentine marriage network (Action, 1993) and the famous Zachary's karate club friendship network (Zachary, 1977). However, networks with relationships among the actors other that friendships can be modeled with graphs too. For example, edges might represent the presence or absence of an item in a group or a temporal relation among the participating entities. As a result, there is no single type of graph that can capture all the characteristics present in all networks. An important distinction surrounding the types of edges of graphs representing social networks are undirected and directed networks. Undirected networks comprise reciprocal relationships, where a relationship from x targetting y implies that the same relationship exists from y targetting x. This is generally true for various online social networks, such as Facebook where there has to be conscent from both parties in order for a relationship to establish. In contrast, directed networks may not have this characteristic. An example of such network is Twitter and the mechanism of followers which does not require conscent to establish a link.

A network is often represented as an illustration of vertices and edges that connect the vertices in the two dimensional plane. Another representation of a graph $\mathcal{G}(N, E)$ is its adjacency matrix $\mathcal{G}_{ij}$ where the value $\mathcal{G}_{ij}$ corresponds to the value (or weight) of the relationship of the edge connecting i and j. Figure 2.1 illustrates a toy example of an

undirected network along with its adjacency matrix. It is worth noting that for this undirected example the adjacency matrix is symmetric around its main diagonal. In contrast, in directed graphs this is not the case where the value of an edge (x, y) might be different than the value of another edge (y, x). The network is also an example of an unweighted graph, where the pairwise relationships among the vertices are binary (presence or absence of relationship).

Certain social networks and social relationships can be represented by special kinds of graphs, such as bipartite graphs that are often used to imprint affiliation networks (Lattanzi & Sivakumar, 2009). Bipartite graphs are graphs whose vertex set can be partitioned into 2 disjoint sets such that no pair of vertices within the same disjoint set are adjacent (occur as endpoints of the same edge). Often, bipartite graphs represent networks whose nodes can be classified into 2 different types of entities. An example would be a network of players and teams where the two types of vertices are connected based on the participation of a player in a team. Bipartite graphs are usually studied using specialized tools and methods that are specifically adjusted to target such networks. Affiliation networks occupy a significant portion of the work presented in this thesis as various networks presented later can be represented as bipartite graphs. Finally, it is worth noting that, similar to the differences among the vertices, networks may comprise edges of different types as well, such as edges that represent friendships and edges that represent categorical existence in the same network. As a result, networks represented by graphs can be heterogeneous, in both the types of vertices and the types of edges.

2.2 Studying Social Networks

Degree and layout One of the most basic measures of studying individual nodes in social graphs is the degree, which is defined as the number of nodes in the neighborhood of a node i. The neighborhood of i is the set of nodes that are adjacent to i and typically represent the friends of i. Similarly to the degree d_i of node i, the *density* of the network is a measure of the fraction of links present in the network and in undirected networks is defined as the average degree normalized by $n - 1$ or the number of edges present over the possible number of edges that could exist:

$$\text{density} = \frac{1}{n-1}\frac{1}{n}\sum_{i=1}^{n} d_i = \frac{|E|}{n(n-1)/2}.$$

Generally, it the context of social network analysis, it is often useful to study the network in terms of its paths among individuals, which is a well studied concept in graph theory. In particular, the shortest paths between vertices can have a physical interpretation with respect to the mechanisms that comprise the network, for example they may convey some form of influence propagation from friend to friend in a social network, or the

spreading of a news story etc. Related to the concept of shortest paths are the ideas of the *average path length* and the *diameter*, where the average path length is the average of the shortest distances between all pairs of nodes while the diameter is the distance between the farthest pair of nodes (the greatest shortest path). The diameter has been studied in empirical networks and analyses suggest that there are cases where, although the density of networks is small, their diameter is surprisingly small too. In contrast, in uncorrelated networks, one would expect the diameter to be inversely correlated with the density of the network.

The diameter of social networks has been studied in the context of *small world networks*, which refers to the idea that larger networks tend to have smaller diameters and average path lengths (Watts & Strogatz, 1998). Specifically, Watts defined the small world network to be a network whose average path length grows proportionally to the logarithm of its order $L = \Theta(n)$. It is worth noting that many empirical networks display signs of the small world effect, such as social influence networks (Kitsak et al., 2010). The process of the creation of small world networks was simulated as a lattice of nodes where each node obtains connections with the nodes that are closer to them and then some of these edges are being rewired to nodes that are farther away.

It is worth noting that the measures and concepts mentioned here can typically only be applied on connected networks, i.e. networks that possess the property that every pair of nodes is connected via a path. In such cases, the analyses are performed either on each maximal connected component individually or in the biggest connected component of the network if that comprises the vast majority of it.

Correlation and clustering Graphs that represent social networks are often much different with respect to random networks in terms of their structure. This structure is what creates certain clustering properties in social networks, which describes the tendency of nodes to be biased with respect to other certain nodes in the network. Networks that exhibit patterns of structure are often called correlated. In contrast, in uncorrelated networks the probability that a node i is connected to a node j is proportional to the product of their degrees ($p_{ij} \propto d_i d_j$).

An example of correlation in a social network might be the tendency of high degree nodes to connect to other high degree nodes, which can often be observed in empirical data. The phenomenon of this tendency is termed positive *assortativity* and has been studied by Newman (M. E. Newman, 2003; M. E. Newman & Park, 2003). Newman highlights that empirical social networks are characterized by positive assortativity. A related concept is *homophily* (Lazarsfeld, Merton, et al., 1954) which is the property via which people have the tendency to connect with other people that are similar to themselves, such as the same gender, race or religion.

In modern day social media and online social networks, phenomena of correlation and assortativity are more prevalent due to the multitude of options given to online users

and the immediacy of their actions. Often, online users decide to act based on their own beliefs and are biased in terms of their online choices, such as the choice of who to follow or what to post. The terms *selective exposure* and *confirmation bias* are coined to describe the users' tendency to expose themselves to information that is already inline with their beliefs or confirms it in the online world. As a result, online users are more likely to connect to other online users that are similar to them in terms of their beliefs on a particular subject.

The aforementioned social phenomena, whose list is not exhaustive, are examples of processes that create correlation in social networks and online social networks. This structure can, in turn, being exploited in order to study and utilize the knowledge inherent in social networks in certain application systems, some of which are given later in Chapter 3 and Chapter 4. One of the simplest ways to extract the clustering properties of a graph is via the local clustering coefficient and the global clustering coefficient, which are measures of the degree of nodes of a graph to cluster together. The local clustering coefficient of a node i is defined as the number of links among vertices in its neighborhood over the possible number of links among them. More formally the local clustering coefficient $C(i)$ of a node i is

$$C(i) = \frac{1}{d_i(d_i - 1)} \sum_{(u,v)} \mathcal{G}_{iu}\mathcal{G}_{iv}\mathcal{G}_{uv},$$

where $\mathcal{G}$ is the adjacency matrix of the graph and (u, v) every pair of vertices in the graph with $u \neq v$. In contrast, the global clustering coefficient is based on triplets of nodes and gives an indication of the clustering properties of the whole graph. It is proportional to the number of triangles in the graph and is defined in the adjacency matrix notation as

$$C = \frac{\sum_{i,j,k} \mathcal{G}_{ij}\mathcal{G}_{jk}\mathcal{G}_{ki}}{\sum_i d_i(d_i - 1)}.$$

Centralities Centrality is a measure that shows how central a node is in the network and is usually a number assigned to each vertex of the graph. A node's centrality might give an indication of how important or influential this node is in the social network and is, therefore, an important measure of individual nodes' position in the network. Several centrality measures have been suggested and, although each measure interprets the concept of importance in a different way, they are generally positively correlated measures. The most basic centrality measure is the degree centrality, which is defined as the normalized degree of a vertex i: $d_i/(n-1)$. Other, more sophisticated centrality measures have been developed, such as the closeness and betweenness centrality. The closeness centrality (Bavelas, 1950) shows how easily a node can reach other nodes and can be defined as the reciprocal of the average distance between i and all other nodes:

$$\text{closeness}(i) = \frac{n-1}{\sum_{j \neq i} l(i,j)},$$

where $l(i, j)$ is the distance of the shortest path between i and j. Finally, the betweenness centrality (Freeman, 1977) shows how well a node is connecting other pairs of nodes and is defined as the fraction of shortest paths that node i lies within:

$$\text{betweenness}(i) = \sum_{s \neq i \neq t} \left(\frac{2\sigma_{st}(i)/\sigma_{st}}{(n-1)(n-2)} \right),$$

where σ_{st} are the number of shortest paths between s and t and $\sigma_{st}(i)$ are the number of shortest paths between s and t that pass through i. The betweenness centrality assumes the value 1 when node i is part of every shortest path of every pair of nodes. The list of centralities mentioned in this section is not exhaustive.

A special class of centrality measures are the eigenvector centrality meausures, whose intuition is based on the assumption that the importance of a node depends on the importance of their neighbors. The most basic eigenvector centrality is defined as the left eigenvector s of the normalized adjacency matrix of the graph $\mathcal{T}$ (the adjacency matrix where each row sums to unity) and is the solution of the following equation:

$$s\mathcal{T} = s.$$

The left eigenvector of a normalized adjacency matrix can also be computed iteratively using the following method:

$$s = \left(\lim_{t \to \infty} \mathcal{T}^t \right)^T \cdot \frac{J_{n,1}}{n}, \tag{2.1}$$

where $J_{n,1}$ is a vector of ones of size n. A popular variant of the eigenvector centrality is Google's PageRank (Page et al., 1999), which is defined as the iterative process based on the following rule:

$$r(t) = d + (1-d)\mathcal{T}^T r(t-1),$$

where $\mathcal{T}^T$ is the transpose adjacency matrix, $r(0)$ is a vector of ones and d a parameter called *damping factor*. The iterative rule is applied consecutively until convergence is achieved. PageRank is a generalization of the eigenvector centrality and is identical to it for $d = 0$, for which the updating rule becomes

$$r(t) = \mathcal{T}^T r(t-1).$$

Upon convergence, and considering Equation 2.1, it is

$$\lim_{t \to \infty} r(t) = \lim_{t \to \infty} \left(\mathcal{T}^T\right)^t r(0) = \left(\lim_{t \to \infty} \mathcal{T}^t \right)^T r(0) = n \cdot s,$$

which demonstrates the equivalence of the PageRank system with the eigenvector centrality for $d = 0$.

Degree distribution and random graphs Following the definition of the degree previously in this section, the degree distribution of a graph $P(k)$ is a discrete probability distribution that determines the fraction of nodes $P(k)$ in a graph that have degree k. The degree distribution is often a basic characteristic of random graph generators. The most famous random graph models are credited to Paul Erdős, Alfréd Rényi and Edgar Gilbert (Erdős & Rényi, 1959; Gilbert, 1959) and are denoted as $\mathcal{G}(n, M)$ and $\mathcal{G}(n, p)$. According to the $\mathcal{G}(n, M)$ model, a graph is chosen uniformly at random from the collection of all graphs that have n vertices and M edges. The $\mathcal{G}(n, p)$ model is closely related and defined by deciding whether to include an individual edge by an independent probability p. The degree distribution of these random models follow the Poisson distribution for large n. A special class of graph generators that result in graphs with degree distribution properties often found in real networks are the scale-free graph generators and are discussed in the following section.

2.3 Scale-Free Networks

Scale-free networks are networks whose degree distribution follows a power law. In particular, the degree distribution of a scale-free graph can be expressed as a probability density function

$$P(k) \propto k^{-\gamma},$$

where $P(k)$ the probability of a vertex to have degree k. The parameter γ is often observed to be between 2 and 3 but occasionaly could be outside this range. A characteristic of power law degree distributions that stem from the above definition is that low degree nodes are far more common than high degree nodes. Such degree distributions are called scale-free because they maintain their shape regardless of the scale at which the independent variable is observed at. In particular, a probability distribution $p(x)$ is said to be scale-free if it satisfies the property

$$p(bx) = g(b)p(x)$$

for any value of b, where $g(b)$ is a function of b. It can be shown that the only distribution that satisfies this property is the power law (M. Newman, 2005).

Scale free networks can often be observed in the real world, for example in the frequencies of words, in biology, in the structure and amount of data transfered in computer networks, in communication networks, the intensity of wars, city populations, various natural phenomena, wealth distribution, in citation networks and others. A comprehensive list of applications of power laws in real world datasets is presented in Clauset et al. (2009). More examples of power laws in real social networks are also found later in this thesis. It is worth noting that, although many real networks have been reported as scale-free networks, this situation has raised controversies (Broido & Clauset, 2019).

Typically power law distributions are presented in logarithmic axes in order to capture their scale-free nature as a straight line. Another common way to plot a power law distribution is with the complementary cumulative distribution

$$P(x) = \int_x^\infty p(x),$$

which is also a power law and, contrary to the $p(x)$ distribution, can be plotted without data binning requirements.

2.4 Preferential Attachment

M. Newman (2005, Section 4) examines the various mechanisms that are responsible for producing power law distributions in natural and man-made structures (Section 2.3). He identifies *preferential attachment* as one of the most convincing and widely applicable mechanisms for generating power laws (Section 4.4) as well as the dominant mechanism that explains the presence of power laws in real networks. In particular, preferential attachment is the phenomenon that often occurs in growing networks, where nodes with more existing connections are more likely to gain connections with new nodes that enter the network in the future. An example might be a scientific paper with a high number of citations, which increases its visibility and makes that paper more likely to be cited in the future. This is possibly the reason why citation networks are observed to have scale-free behavior (Price, 1963). The preferential attachment mechanism was previously known as the *Gibrat* principle, the *Yule process*, the *Matthew effect*, or *cumulative advantage*.

The most widely known type of graph generator that formulated the modern version of the preferential attachment mechanism is described by Barabási and Albert (Barabási & Albert, 1999; Albert & Barabási, 2002). In defines two conditions for the graph generator to result in scale-free graphs:

1. Growing scheme, where newborn nodes enter the network one by one and connect with m different older vertices.

2. The probability of a newborn node to connect with an existing node is proportional to the degree of the existing node.

In their works, Barabási and Albert provided analytical proofs (Albert & Barabási, 2002, Section VII.B) that the described model results in a power law degree distribution. This property was initially observed by Yule (Yule, 1925) and later named the *Yule–Simon discrete probability distribution*:

$$f(k;\rho) = \frac{\rho\,\rho!\,(k-1)!}{(k+\rho)!}.$$

Albert and Barabási indirectly showed that for their model $\rho = 2$, thus

$$f(k) \propto \frac{1}{k(k+1)(k+2)},$$

which for sufficiently large k (tail of the distribution) is a power law

$$f(k) \propto k^{-3}.$$

2.5 Utilizing Social Networks

The behavior and the actions of users in social networks defines the information inherent in these networks. Each individual social network, whether a human community or an online social network, is subject to the users' manipulation and most of the information embedded in it is a direct result of the users' actions. In particular, each user, driven by their own motives, beliefs and criteria, makes the corresponding choices in the network, which might be the decision to follow someone on Twitter, or a post on Facebook or as simple as a *like* (or *favorite*) action. Collectively, this type of behavior and actions is what defines the information inherent in the corresponding social network, and ultimately stems from the criteria and decisions of the entities participating in it.

While the information referring to individual users is usually easier to obtain, our goal is to utilize the social network as a source of information for a third system. Therefore, the information we seek to extract doesn't refer to individual users but reflects the average behavior of the users in a network about a particular subject of a distribution of their behavior. This type of information is usually implicit and not in a form readily available to utilize; methods of derivation of information need to be applied in order to extract the required knowledge. This knowledge is, in turn, exploited to power or enhance another system that relies of the behavior of the users for its effectiveness. While there are numerous ways of achieving this goal, in this section we present the methods that are commonly used and mentioned in this thesis.

2.5.1 Graph Projections

One of the most used methods of utilizing social networks in the work described in this thesis is the *graph projection*. A projection of a graph is a transformation that assigns a value to each pair of vertices in the graph and is often related to the concept of *link prediction* (Liben-Nowell & Kleinberg, 2007; Zhou et al., 2009) or *vertex proximity* (Goyal & Ferrara, 2018). Graph projections are often used as a preliminary stage of a methodology, either because they can compress a large graph into a more manageable size or because they can convey information about the pairwise relations of the nodes that would otherwise not be available.

Graph projections are commonly used on bipartite graphs or affiliation networks because their operation fits the description of these networks perfectly. Specifically, in bipartite graphs, relations among the nodes of one of the disjoint sets are implicit as there are no edges among these vertices; a graph projection is then able to capture these implicit relations and quantify them. This type of projection is called the *one-mode projection* of the bipartite graph and it is a very useful method of assigning pairwise weights on the relations among the vertices of one of the disjoint sets of the network. These relations can, in turn, portray a measure with physical interpretation or can be manipulated to have one. In most cases, for the work presented in this thesis, these measures convey a form of similarity among the participating entities.

In particular, the one-mode projection of a bipartite graph $\mathcal{G}(A, B, E)$ –consisting of the disjoint sets A and B whose edge set is E– onto A is defined as a unipartite graph $\mathcal{G}'(A, E')$, such that every pair of nodes (u, v) is linked with a weighted edge of weight $\beta(u, v)$. Because graph projections transform the original graph and, therefore, incur information loss over it, the choice of the weighting method β is important. While there are many methods of weighting the pairwise relations in a projection, each one may interpret this relation in a different way and, hence, the decision over which method to use is a subject of study of individual problems. We do not further discuss the concept of projection here. Instead, it is more thoroughly discussed in individual problems of this thesis in Sections 3 and 4.

2.5.2 Community Detection

One of the most common perspectives of structure inherent in social networks is attributed to the tendency of vertices to cluster together. Unlike random networks, the edge distribution in real social networks is not homogenous but has underlying structure. This phenomenon has been described as *community structure* (Girvan & Newman, 2002), which describes the property of networks to form natural groups of nodes based on some criteria. Therefore, the most innate way to study this property is through the analysis obtained via the various methods and techniques of *graph clustering* or *community detection*.

The aim of community detection is the uncover the underlying structure in a social network by identifying the modules of vertices that best describe this structure. Community detection has seen extensive use in social network analysis and particularly in social media. Some applications are given in Papadopoulos et al. (2012) and include topic detection, tag disambiguation, user profiling, photo clustering, event detection and others. Community detection has also seen use on other topics, such as criminology, public health, politics and community evolution prediction. More applications are given in Karataş and Şahin (2018) and the references therein. In the work described in this thesis, community detection is being applied to determine the political orientation of particular Twitter users in Section 3.

Community detection has not received a single definition throughout the literature but in the context of social networks its general concept is to partition the nodes of the network in such a way such that on average two nodes that belong to the same cluster display higher or much higher similarity than two nodes that belong to different clusters. It is also worth mentioning that this partition may be overlapping (cover). The intuition behind community detection on social graphs is that it will reveal the natural groups of vertices that inherently exist in the network such that various arguments can be stated about the structure, the properties and characteristics of that network that relate to the physical world.

Many community detection approaches have been developed and exist in the literature. Perhaps the most elementary type of graph clustering are the *divisive algorithms*; one of the most popular is the algorithm of Girvan and Newman (Girvan & Newman, 2002). According to the algorithm, edges are initially assigned a meausure of importance, called the betweenness edge centrality and then the highest centrality edge is removed. This operation may split the network into two connected components. The process is repeated until the network reaches the desired state, i.e. the desired number of components, which are considered the clusters.

Other types of clustering methods include the *hierarchical methods* and the *partitional clustering*. Hierarchical clustering is a generalization of the divisive method and also include the *agglomerative approach*, according to which the community structure starts considering each individual vertex as a singleton community and then iteratively merges communities that are very similar to each other according to a similarity function. The partitional clustering also depends on a similarity function among the vertices of the network and the goal is to find the centers of the clusters using a measure that relies on the similarity function, for example a quantity that tries to spread the centers throughout the implicit two-dimensional geometry of the graph. Common types of partitioning algorithms include the k-clustering, the k-center and the k-median approaches.

A large category of clustering methods depend on the optimization of the meausure called the *modularity*. More formally, given a graph $\mathcal{G}(N, E)$ and a partition of the graph, the modularity quality function is equal to

$$Q = \frac{1}{2|E|} \sum_{ij} \left(\mathcal{G}_{ij} - P_{ij} \right) \delta(i, j),$$

where $\mathcal{G}_{ij}$ is the binary indicator of whether an edge between i and j exists, $\delta(i, j)$ is a binary value indicating whether i and j are on the same cluster on the partition, and P_{ij} represents the probability of i and j sharing a common edge in the *null model*. The null model is a graph which retains some of the properties of the original graph but without community structure and usually comprises the random graph with the same degree sequence as the original. Therefore, the quantity P_{ij} becomes

$$P_{ij} = \frac{d_i d_j}{2|E|}.$$

The intuition behind the modularity function is to estimate the difference of the social network with the null and imprint the difference between the presence or absence of an edge with its expectation in the null model ($\mathcal{G}_{ij} - P_{ij}$). Thus, the modularity-based algorithms aim at finding an appropriate partition (the values of δ) such that the modularity quantity is maximized. In a sense, it is a way of saying that whether two nodes should be included in the same group does not depend on their similarity but on the difference of their similarity from the expected similarity from the null model. The most famous modularity maximization algorithm is known as *Louvain optimization* and is a greedy method that has seen extensive use in the field of social networks (Lancichinetti & Fortunato, 2009).

Special types of graph clustering methods have been developed to accomodate the needs of special types of graphs, for example affiliation networks. In such cases, the similarities between vertices are only implicit and can be uncovered using a transformation, such as the projection methods discussed earlier. Ragardless of the type of graph, the projections are useful for graph clustering because they convey the similarities among the vertices, on which many clustering algorithms rely upon. More categories and details about clustering and community detection methods can also be found in the surveys of Fortunato (2010) and Schaeffer (2007).

2.5.3 Graph Layout

A large class of algorithms for studying social networks, the network layout, is the transformation of the vertices of the network in the m-dimensional space such that some of the properties inherent in the network are conveyed in the resulting illustration. The most common graph layout algorithms are the network visualizations which are very common illustrations of graphs in this thesis. In this section, we also discuss the problem of minimum linear arrangement which, despite not usually discussed in the context of graph layouts, displays similarities to layouts in terms of the positioning of the vertices and it has not seen significant utilization in the context of social networks.

Force Directed Drawing

One of the most natural ways of the representation of the layout of a graph is via its embedding in the two dimensional plane. In general, the problem in such layout is to find a function of a node that returns a two dimensional vector of its representation such that an overall quality function for the entire graph is optimized. Graph visualization is an important tool for the analysis of networks as it can demonstrate some of its natural properties that are difficult to imprint in the raw data. According to Newman (M. E. J. Newman, 2010):

> Visualization can be an extraordinarily useful tool in the analysis of network data, allowing one to see instantly important structural features of a network

> that would otherwise be difficult to pick out of the raw data. The human eye is enormously gifted at picking out patterns, and visualizations allow us to put this gift to work on our network problems.

As graph visualization is a general problem, there exist many methods of transforming a network into a 2D representation and each one of them interprets the position via the optimization of different quality functions. There are, for example, circular layouts (Doğrusöz et al., 1997) where nodes are placed on a circle and rearranged such that adjacent nodes are positioned nearby, spectral layouts that rely on the eigenvectors of the Laplacian matrix of graphs (Koren, 2005), and others. It is worth mentioning that, although two-dimensional drawings are the most common due to their natural visualization in printed and reading media, often visualization methods can be generalized for higher number of dimensions.

One of the most frequently used visualization method is the *spring layout* or *force directed layout*, which consists of a class of algorithms for the graph visualization in (usually) the 2D space, such that it highlights the underlying structure of the network (Kobourov, 2012). Force directed visualization algorithms employ natural phenomena to simulate the position of nodes. In the pioneering work of Eades (1984), each node is assigned an electric charge, such that all nodes have the tendency to repel each other due to Coulomb's law. At the same time, vertices that are linked via an edge in the original network are connected with the mechanical equivalent of a spring that pulls them closer together according to Hooke's law. The system is allowed to run for a number of rounds until convergence is achieved. The result of this simulation highlights the clustering features of the data, because groups of nodes that are tighly packed with edges in the network have the tendency to form geometric clusters in the force directed visualization of the network. The close association between modularity clustering and force directed placement has also been demonstrated before (Noack, 2009).

The relation of modularity clustering and force-directed layout can also be illustrated in the example of Figure 2.2. The figure displays a force-directed visualization of the unweighted Karate Club network, where vertices are colored based on their modularity class on the optimal 2-partition of the network. The optimal partition of the vertices into 2 groups has been found using a brute-force method for all 2-partitions to be approximately equal to 0.3718. The illustration shows the instructor (node 1) and the president (node 34) take the roles of the centers of their respective groups. The vertices are sized on a linear scale by their degree to illustrate their importance. The figure also includes a separator curve between the clusters for grayscale media.

The Minimum Linear Arrangement Problem

The *Minimum Linear Arrangement* (MinLA) problem is another graph layout problem and aims at placing the vertices of a graph in a one-dimensional arrangement, such that the MinLA cost is minimized. In particular, it consists in finding an ordering of the

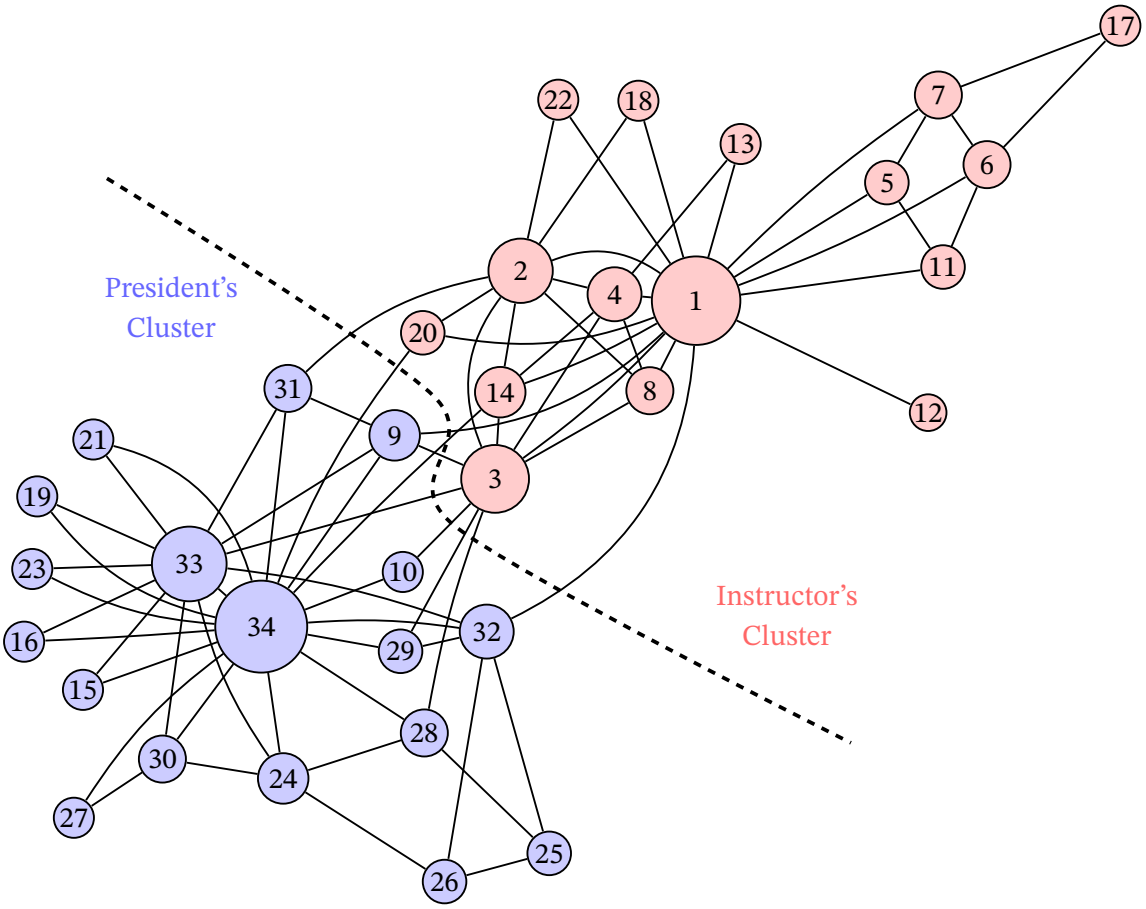


Figure 2.2: A force-directed visualization of the Karate Club network, where nodes are colored by their modularity class for the optimal 2-partition of the vertex set. Vertex 1 stands for the instructor and node 34 for the club president. The nodes have been sized on a linear scale based on their degree. The figures also includes a separation curve between the clusters for grayscale media.

nodes of a weighted graph, such that sum of the weights of its edges is minimized. More formally, given a finite graph $\mathcal{G} = (V, E)$ of order n with weighted adjacency matrix w, the MinLA problem is the problem of finding a vertex labeling $f \rightarrow \{1, 2, ..., n\}$ such that the sum

$$\sum_{(u,v)\in E} w_{uv}|f(u) - f(v)|$$

is minimized over all possible labelings (Safro et al., 2006). In Chierichetti et al. (2009) additional minimum arrangement problems are being defined with different cost functions, in particular the *minimum logarithmic arrangement* and the *minimum gap logarithmic arrangement* that are modifications of the MinLA problem and similar in nature to it.

In contrast to a graph visualization, the result of the application of the MinLA problem is rarely useful as a way to observe the geometry of a graph. While it can highlight the clustering features of a network –we will later show this in Section 3–, there is no notion of distance as all nodes are placed in the integer range $[1, n]$. However, it has been applied to various scientific fields, for example in VLSI design (Petit, 2003) in order to minimize the electrical resistance of a circuit, or in a theoretical level to study the compressibility of large networks (Chierichetti et al., 2009).

Regarding its computational complexity, on general graphs MinLA is an NP-complete problem, thus one has to resort to heuristics or approximation algorithms to obtain a solution. However, there are no "good" approximation guarantees for the MinLA problem, either. The best known result, is the $\mathcal{O}(\sqrt{\log n} \log \log n)$-approximation algorithm presented in Feige and Lee (2007). On the other hand, in Ambuhl et al. (2007) it is shown that no Polynomial Time Approximation Scheme (PTAS) exists for MinLA and in Raghavendra et al. (2012) that it is SSE-hard to approximate MinLA to any fixed constant factor.

2.5.4 The DeGroot Model

Degroot (1974) presented a simple yet efficient model about opinion diffusion in a social graph. The core idea of his model is that individuals tend to adopt the opinions of their friends. According to the model's *opinion update rule*, given a social graph $\mathcal{G} = (V, E, O)$, where V represents the vertices (i.e., individuals), E the edges amongst them (i.e., friendships) and O the opinions of nodes $i \in V$ about a specific topic as real valued o_i, each individual i updates o_i to o_i' by averaging the opinions of its friends. When *trust factors* are introduced to the friendships (i.e., weights), each member updates its opinion according to the weighted average of its friends' opinions. The process is repeated and, under certain condition, the opinions of the nodes converge signifying a consensus in the graph. DeGroot underlined the mathematical coherence of the process to Markov chains. He proved that the final opinion, when convergence occurs, depends solely on the structure of the graph and the initial opinions of its members.

In Ghaderi and Srikant (2013), Ghaderi and Srikant enriched DeGroot's model with *stubborn agents* (i.e. nodes that are fully or partially biased towards an opinion) and studied its convergence. They remarked the common underpinnings of their extension with Markov chains and proved that "the model converges to a unique equilibrium where the opinion of each agent is a convex combination of the initial opinions of the stubborn agents". Moreover, the contribution of stubborn agent s in node's i final opinion is the probability of a random walk hitting s given it started from i, namely the *hitting probability*.

The conditions of convergence in the DeGroot model have been studied by Golub and Jackson (2010). The analysis presented therein include the instances of the enrichment of the model by Ghaderi and Srikant in the presence of stubborn nodes. The authors formulate the theorem of convergence of the DeGroot model:

> The process is convergent if and only if every set of nodes that is strongly connected and closed is aperiodic.

More formally, a strongly connected set of nodes is a set of nodes, for which there exists a path among each pair of vertices and a closed set of nodes is the set in which there exists no outgoing edge between any node in the set and any node outside the set. A maximal strongly connected and closed component can be seen as a stubborn component, i.e. a group of vertices that are biased towards an opinion. Finally, an aperiodic graph (or subgraph) is that for which the greatest common divisor of the lengths of all cycles is unity.

2.5.5 Recommendation Systems

Recommender systems (or recommendation systems) are tools that provide suggestions that are of interest to a user (Ricci et al., 2011). These systems typically attempt to predict the rating or preference of a user towards a particular item of interest. This process is usually part of our nature, for example by asking our social circles what they think about items of interest, opinions that are spread via word of mouth, recommendation letters, reading online reviews and others (Resnick & Varian, 1997). Recommendation systems generally aim to provide an automatic way to perform these tasks with minimal or no user intervention.

Recommendation systems are a broad topic and can be categorized based on their methodology and their goal. The most common methods of recommendation systems are content based, link analysis based, and collaborative filtering. Content based systems usually match the interests of users based on their own profile individually and not on information generated by other users. In contrast, in collaborative filtering approaches, the recommendation systems are allowed to quantify the similarities among users and make recommendations based on the preferences of similar users. Finally, link analysis based methods extract information relevant to the recommender from the structure

of complex networks. The goal of the system is another criterion of categorization among the recommendation systems as well. Common uses of these systems are found in social networks for user recommendation (or friend recommendations), activity recommendations, and other online platform recommendations, for example in movies, shopping and others. Comprehensive surveys on recommendation systems are found in Bao et al. (2015) and Eirinaki et al. (2018).

Online social networks are a special kind of knowledge source for modern recommendation systems due to the volume of information that is generated on a daily basis as well as the interactivity of the platforms' users with the platform itself or with other users. In particular, the emergence of location based social networks (LBSNs) have given rise to location recommendation systems, i.e. recommenders that quantify the preference of a user to particular locations or points of interest (POIs). LBSNs have the unique characteristic that they connect users with locations and other rich-content information about them. Zheng and Zhou (2011b) describe LBSNs as:

> A location-based social network does not only mean adding a location to an existing social network so that people in the social structure can share location embedded information, but also consists of the new social structure made up of individuals connected by the interdependency derived from their locations in the physical world as well as their location-tagged media content, such as photos, video, and texts. Here, the physical location consists of the instant location of an individual at a given timestamp and the location history that an individual has accumulated in a certain period. Further, the interdependency includes not only that two persons co-occur in the same physical location or share similar location histories but also the knowledge, e.g., common interests, behavior, and activities, inferred from an individual's location (history) and location-tagged data.

Location based social networks and their role in recommendation systems is being discussed in Section 4, where novel information encoded within the structure of such networks is being utilized to support a POI recommendation system.

2.6 Random Sampling

Random sampling refers to the problem of collecting a subset of items (sample) from a larger set of elements (population) in which the items of the sample are chosen randomly. More formally, given a population of items $\mathcal{P} = \{p_1, p_2, \dots, p_n\}$ and a list of possible samples $\mathcal{S} = \{s_1, s_2, \dots, s_{|\mathcal{S}|}\}$, a random sampling method is a function that assigns a probability $\Pr(s_i)$ to each of the subsets in $\mathcal{S}$, such that some $\Pr(s_i)$ may be zero and

$$\sum_{i=1}^{|\mathcal{S}|} \Pr(s_i) = 1.$$

It can be shown that random sampling is a fundamental mechanism on social processes that explain certain phenomena and is presented here as an introduction to the terminology and notation of the problem. We note that the definitions and abbreviations found throughout this thesis are commonly found in the literature (Hanif & Brewer, 1980; Rosén, 1997b; Tillé, 2006).

There exists two major categories of random sampling: with replacement and without replacement. For the purposes of this thesis, the problem of random sampling without replacement is more relevant. As a result, the notation and descriptions througout the text refer to random sampling without replacement. In this case, each sample s_i is a set that doesn't contain multiple items. We note that it is the final result that determines the classification of a random sampling design as with or without replacement, and as such there exist random sampling designs without replacement with an intermediate step of sampling with replacement.

2.6.1 Inclusion Probability

A fundamental concept of random sampling designs is the *inclusion probability* of elements of the population or groups of elements of the population. In the case of individual elements, the inclusion probability is referred to as the first order inclusion probability and is equal to the probability of a particular item to exist in the final selected sample. More formally, the (first order) inclusion probability of item p_r is

$$\pi_{p_r} = \sum_{i=1}^{|\mathcal{S}|} \Pr(s_i) \cdot I(p_r; s_i),$$

where $I(p_i; s_j)$ is a binary variable indicating the presence (1) or absence (0) of the item p_i in the sample s_j.

Equivalently, the higher order inclusion probability of groups of elements can be defined. We note that every possible group of elements can be described using the sample notation s_i, even if the respective probability of selection is zero ($\Pr(s_i) = 0$). The k-order inclusion probability of the k elements contained within s_r is the probability that the final selected sample contains all the items of s_r and is defined to be

$$\pi_{s_r} = \sum_{i=1}^{|\mathcal{S}|} \Pr(s_i) \cdot I(s_r; s_i),$$

where $I(s_r, s_i)$ is a binary variable indicating the presence (1) or absence (0) of all the elements of s_r in s_i. It is also easy to see that $\pi_{s_r} \geq \Pr(s_r)$.

2.6.2 Higher Order Inclusion Probability: An Example

Assume a population of 4 elements $\mathcal{P} = \{A, B, C, D\}$ and 2 different random sampling designs without replacement that randomly select a sample of 2 elements from $\mathcal{P}$: RS-A

Pair	RS-A	RS-B
AB	1/4	1/3
AC	1/4	0
AD	0	1/6
BC	0	1/6
BD	1/4	0
CD	1/4	1/3

Table 2.1: The second order inclusion probabilities of all pairs of elements for the RS-A and RS-B random sampling designs.

and RS-B. The sample space of RS-A is AB, AC, BD, CD with selection probabilities 1/4 each and the sample space of RS-B is AB, CD, AD, BC with selection probabilities 1/3, 1/3, 1/6 and 1/6 respectively. Here, all selected samples are of size 2.

For the RS-A design, the first order inclusion probability of A is the sum of the selection probabilities of the AB and the AC samples, which are the only ones that contain the element A:

$$[\text{RS-A}] \quad \pi_A = \Pr(AB) + \Pr(AC) = 0.5.$$

Similarly, it can be shown that the inclusion probabilities of all elements for both random sampling designs are 0.5. While RS-A and RS-B do not differentiate in terms of their first order inclusion probabilities, this is not the case for the second order inclusion probabilities, for example for the BD pair:

$$\begin{aligned} [\text{RS-A}] \quad & \pi_{BD} = \Pr(BD) = 1/4 \\ [\text{RS-B}] \quad & \pi_{BD} = 0. \end{aligned}$$

Table 2.1 shows the second order inclusion probabilities of all pairs of elements for both random sampling designs.

2.6.3 Weighted Random Sampling

An important class of random sampling designs is *weighted random sampling* (WRS) (Efraimidis & Spirakis, 2006) or *unequal probability random sampling* (Tillé, 2006) or *varying probability random sampling* (Arnab, 2017), where each element in the population p_i is also assigned a parameter or weight x_i, that determines its (first order) inclusion probability. In contrast, in equal probability random sampling, the first order inclusion probabilities of all elements in the population are equal as demonstrated in the example in Section 2.6.2.

While in unweighted random sampling the weights of the items are missing and, hence, there is only a single way to compose the first order inclusion probabilities, in weighted random sampling there is an arbitrarily large number of ways that the parameters x_i can be interpreted to compose the first order inclusion probability combinations. As a result, each weighted random sampling design interprets the given weights in a different way which leads to differences in the respective inclusion probabilities.

Table 2.2: An analytical unequal probability example of n elements with $x_1 = 2$ and $x_i = 1, i = 2, \dots, n$. The table displays the inclusion probability of the heavy node (node 1) as the π_1 column and the difference among the sampling designs.

Design	π_1	strπps diff.
strπps	$\frac{4}{n+1}$	0
Draw-by-draw	$\frac{2}{n+1} + \frac{n-1}{n+1} \cdot \frac{2}{n}$	$\frac{2}{n^2+n}$
Conditional Poisson	$\frac{4n-4}{n^2-n+2}$	$\frac{4n-12}{n^3+n+2}$

A rigorous survey of weighted random sampling methods was given in Hanif and Brewer (1980) and Brewer and Hanif (1983); the reader may refer to these resources for more detailed explanation of the different weighted random sampling designs and their classification on categories based on different criteria. Other unequal probability random sampling designs are given in Tillé (2006), Y. G. Berger and Tillé (2009) and Grafström (2010).

Three of the most utilized WRS designs that are relevant in the concepts described in this thesis are:

1. The conditional Poisson design (Hajek, 1964).
2. The draw-by-draw selection (Yates & Grundy, 1953).
3. The strπps scheme.

These WRS designs are without replacement and with constant sample size m. According to the conditional Poisson design, each individual element i of the population is included in the sample with an *independent probability* proportional to x_i. If the sample is not of size m, it is rejected and the process is repeated. According to the draw-by-draw selection of Yates-Grundy, the elements of the sample are selected one by one with *selection probability* proportional to their weights. If a selected element has been seen previously, it is rejected and the selection round is repeated. The strπps (inclusion probability strictly proportional to size) scheme is a special case of WRS where the weights coincide precisely with the *inclusion probabilities* of the elements of the population ($\pi_i \sim x_i$). This implies that all strπps designs are equivalent in terms of the first order inclusion probabilities of the population elements. The strπps design is also known as the *ratio estimator property* (Brewer & Hanif, 1983, Section 1.4). The cases of the draw-by-draw and strπps design are also discussed in Efraimidis (2015). It is worth noting that the distinctions are not mutually exclusive; for example a strπps design may operate in a way that resembles the draw-by-draw method of Yates-Grundy.

The differences regarding the first order inclusion probabilities of these designs are demonstrated in a simple example in Table 2.2. The table shows an analytical example of n elements with sizes $x_1 = 2$ and $x_i = 1, i = 2, \dots, n$, i.e. one heavy element with double the size of the other $n - 1$ elements. The analysis shows that, in at least one

setting, the random sampling designs are different as they lead to different inclusion probabilities for finite n. The table also displays the quantified differences among the designs via which it can easily be shown that they converge asymptotically for large n. It is worth mentioning that for single item samples, these WRS designs are identical.

A special class of random sampling algorithms are the reservoir sampling algorithms, which are memory efficient and can work under arbitrarily large populations. One such algorithm is mentioned in Efraimidis and Spirakis (2006), for which the parameters are being interpreted as the selection probabilities and is equivalent to the draw by draw selection as proven in K.-H. Li (1994a), and another one in Chao (1982), which is a strπps sampling scheme. Naturally, some existing weighted random sampling methods that are not published in a scheme compatible with reservoir sampling, can be transformed into one pass algorithms using reservoirs. For example, the sequential Poisson can be implemented with a reservoir by assigning one random variate X_i in $(0, 1)$ for every element and selecting the elements with the smallest values of X_i/p_i (Ohlsson, 1998, Section 2.2), using the algorithm described in Efraimidis and Spirakis (2006) with a priority queue. The Pareto design can also be implemented using a reservoir using the same principle by selecting the elements with the smallest values of $X_i(1-X_i)/p_i(1-p_i)$ (Rosén, 1997b, Section 4.3).

Chapter 3

Application: Deriving the Political Affinity of Twitter Users

In this chapter, we show that the structural features of the Twitter online social network can divulge valuable information about the political affinity of the participating nodes. More precisely, we show that Twitter followers can be used to predict the political affinity of prominent Nodes of Interest (NOIs) they opt to follow. We utilize a series of purely structure-based algorithmic approaches, such as modularity clustering, the minimum linear arrangement (MinLA) problem and the DeGroot opinion update model in order to reveal diverse aspects of the NOIs' political profile. Our methods are applied to a dataset containing the Twitter accounts of the members of the Greek Parliament as well as an enriched dataset that additionally contains popular news sources. The results confirm the viability of our approach and provide evidence that the political affinity of NOIs can be determined with high accuracy via the Twitter follower network. Moreover, the outcome of an independently performed expert study about the offline political scene confirms the effectiveness of our methods. The results presented in this chapter are published in Stamatelatos et al. (2018) and Stamatelatos et al. (2020).

3.1 Introduction

Twitter, an online news and social networking service, has been subject of scientific research for at least a decade. Users in Twitter can follow other users in order to receive short messages posted by them, which are called tweets. The follower relationships of Twitter naturally convey an inherent directed graph structure, where vertices are the user accounts and edges represent the follower-to-followee relationship. The interpretation of these links varies across contexts: they may represent intimate relationships, common interests, an intent in news briefing and many others.

The significance of Twitter in research is partially because it supplies means to com-

prehend social relationships and influence dynamics in human societies. Various studies examine the structural and topological characteristics of the Twitter network, for example via concepts related to user influence and centrality (Riquelme & González-Cantergiani, 2016) while others focus on extracting information from the content of tweets (Giachanou & Crestani, 2016). Furthermore, the Twitter network has been previously used for a multitude of practical applications, for example stock market predictions (Nisar & Yeung, 2018), event detection (Hasan et al., 2018) and geo-locating users (Dredze et al., 2016).

A distinct characteristic of Twitter is the presence of politically related actors, for example politicians or other party representatives, public officials, candidates as well as news media. These actors engage on the social platform as part of their political campaigns or utilize it as a means of political deliberation and advocacy (Parmelee, 2014). The topic of political deliberation in social networks is relevant and has received substantial attention, especially through its applications to the identification of political bias in news sources.

In this work, we study the topic of deriving the political affinity of particular nodes of interest (NOIs) by using the structural features of the Twitter network. More specifically, we consider the NOIs to be the members of the Greek Parliament (MPs) and the most popular news sources, although the set of NOIs can be enriched with other politically engaged actors as well. Our approach focuses on two primary objectives. The first objective is to confirm that Twitter links between the NOIs and their followers can be used to identify the political affinity of the MPs and establish suitable methods to accomplish this. Our findings indicate that the Twitter follower network can portray with very high precision the affinity of political actors. Our second objective is to extend the application of this methodology to determine the political affinity of the most popular news sources, a natural extension given the strong relationships among political actors and news media. We argue that the results are promising and in agreement with the actual political scene, although not as consistent as the findings of the first objective.

Since, however, there is no single interpretation of political affinity, we determine three different perspectives and establish analytical methods that comply with each one. The *group affiliation* refers to the identification of groups or clusters of NOIs with the same or similar affinity. The *bipolar arrangement* projects the NOIs in a one dimensional arrangement in respect to a relevant measure, for example the left-to-right political axis. Finally, the *influence factors* constitutes a way to quantify the affinity of each NOI relative to another entity, for example the political parties.

The methods we propose in this work rely only on the social ties formed among relevant parties in the network and don't require any prior knowledge regarding the political standing of the involved entities. Our approach utilizes the nodes of the implicit graph structure simply as their Twitter IDs, and no additional knowledge about these

user accounts is required. Furthermore, our methods are easily reproducible and can be implemented without complex filtering or preprocessing. They attain very high accuracy, even on a complex political scene with a large number of political parties. An important feature of our methodology is that it leverages the knowledge of the network to determine the political standing of a NOI. This is in direct contrast with utilizing the NOI's own explicit profile, for example tweets or friends, which portrays what the NOI is trying to convey to the network, rather than the opinion of the network about them. An additional effect is that users of importance cannot easily handpick their followers, who adapt to the online and offline political scene.

Studying the topic of deriving the political affinity of news sources is appealing because it constitutes a primitive technique of inducing higher level knowledge from public information. In particular, results about the political affinity of news sources can potentially characterize the news media scene of a country as a whole, for example if it is biased or it favors only part of the political spectrum. Moreover, another application might be the identification of political bias in particular news articles or even the classification of fake news.

Overall, our approach relies on the assumption that people's political preferences will, on average, reflect those of the politicians or the news sources they follow, a phenomenon described as *selective exposure*. Prior literature on this topic suggests this assumption is reasonable since people seek after information from those with similar political views (Garrett, 2009). In the context of our study, the interpretation of selective exposure dictates that the following decisions of Twitter users provide information about their own perceptions of both their ideological position and that of the political accounts they follow (Barberá, 2015). Previous research demonstrates that the assumptions that our approach is relied upon are well founded in a news media framework as well. For example, in Gentzkow and Shapiro (2010) it is stated that readers have an economically significant preference for like-minded news, which is consistent with our assertions.

Finally, in this chapter, we make use of the *overlap coefficient*, a measure of similarity between two sets, in particular the follower sets of the NOIs. This measure appears for example in Borgatti and Halgin (2014) for the purpose of studying affiliation networks and in Vijaymeena and Kavitha (2016) for text mining applications. To the best of our knowledge, the overlap coefficient has not seen extensive use until now in the context of social network analysis.

The highlights of our contribution are summarized as follows:

- Further proof of the selective exposure phenomenon, targeted for the Twitter network, as well as additional evidence that followers can portray the political leaning of their followees.
- The analytical formulation of three distinct perspectives of political affinity (the

group affiliation, the bipolar arrangement and the influence factors) and the suggestion of techniques suitable for each perspective.

- A structural dataset acquired via the Twitter API comprising the nodes of important political influence in Greece along with their follower sets.
- The application of novel techniques, specifically the Minimum Linear Arrangement problem, which is not mentioned in the Social Network Analysis literature.
- The promotion of the overlap coefficient as a measure of pairwise similarity.

The chapter is organized as follows. In Section 3.2 we explore the recent literature in respect to political concepts in social networks. The dataset used in this work as well as the methodology are described in Section 3.3. In Section 3.4 we demonstrate the existence of rich political information within the Twitter follower dataset by evaluating the effectiveness of our methods and, moreover, lay out the experimentation settings. In Section 3.5, our methodology is applied on the combined MP and news sources dataset to assess the political affiliation and orientation of the news sources. The results are evaluated against the replies of an expert survey that was conducted for this purpose. Finally, Section 3.6 concludes this chapter and presents suggestions for future work.

3.2 Related Work

Previous research demonstrates that the concept of social network analysis in Twitter and other online social networks is a very active field. In Zarrinkalam et al. (2018) the authors build interest profiles of social network users based on the homophily principle; users tend to interact with users with common interests or preferences. The review in Riquelme and González-Cantergiani (2016) summarizes methods of quantifying the influence and popularity of users, targeted at the Twitter network, while in Celik and Dokuz (2018) the similarity among users in social communities is detected based on the similarities of their spatial history profiles. In this study we focus exclusively on the political aspect of social interactions while our objective is to identify the political interests of influential users based on their followers.

A number of previous studies have promoted concepts related to the detection and analysis of political affinity. In Tumasjan et al. (2010), the authors show that Twitter is used extensively for political deliberation and evaluate whether tweets reflect the current offline political sentiment. In Pennacchiotti and Popescu (2011), the values of user attributes such as political orientation or ethnicity are inferred, while in Maynard and Funk (2012) an example application to determine political leanings from tweets is demonstrated. These methods operate by examining the content of tweets while the approach presented in this chapter utilizes algorithms that only rely on the topological and structural characteristics of the Twitter network.

Furthermore, in Verweij (2012), the authors construct the politician-journalist graph and attain multiple conclusions regarding the network structure. Moreover, the study of Boutet et al. (2013) is the identification of the characteristics of political parties and the political leaning of users in social media. The data scheme used in these reports is similar to the one used here but our focus and methodology are distinct.

Three studies that share common characteristics with our political affinity perspectives are Conover et al. (2011), An et al. (2012) and Le et al. (2017). These works investigate political information within social networks but each from a different perspective. In particular, the authors of Conover et al. (2011) employ clustering methods in respect to the left and right leaning tweets, a concept that is related to our clustering approach. The authors of An et al. (2012) propose a methodology for positioning news media on a one-dimensional Euclidean political space via the Jaccard similarity of their follower sets. This format is in accordance with the scheme produced by the application of the minimum linear arrangement problem in this work. Similarly, in Le et al. (2017), the political slant of articles are evaluated through the projection of the journalists' political preferences. The methodology presented is able to quantify the slant of a news article in a scale of -1 to 1, which can be parallelized with the quantifiable measures from the DeGroot model application of this study.

The Greek political scene was previously studied in Tsakalidis et al. (2018), in which the authors employ a learning model to predict the voting intentions during the 2015 Greek bailout referendum. Relevant tweets dating before and after the referendum are leveraged in order to examine the intentions of this spontaneous in nature event. In the context of referendums, the Twitter network is utilized in Marozzo and Bessi (2017) as well to study the effects of news media in the 2016 constitutional referendum in Italy. The potential of Twitter as a platform of information dissemination and dialogue in Greece is also examined in Poulakidakos and Veneti (2019) by applying content and thematic analysis on the tweets of the two biggest Greek political parties.

While our case study is the Greek political scene, previous literature on the behavior of political actors in the USA is very common. In Sainudiin et al. (2019), the authors examine the Twitter linkages between five major American political leaders, among them US President Donald Trump, with eight America hate groups (e.g. Anti-Immigrant and White-Nationalist). This appears to be in parallel with our investigation of linkages among politicians and news media. The follower-followee connections of the Twitter network are also utilized in (King et al., 2016) to identify a latent ideological dimension concerning political actors in USA's political scene.

An interesting study in Golbeck and Hansen (2014) attempts to infer the political leaning of news outlets in the US by characterizing the followers and then relaying the followers' preference to the news outlets that they opt to follow. The authors claim that, overall, users tend to follow politicians with similar views and that those who follow Congresspeople on Twitter may have more polarized political tendencies that the

overall US population. The results are achieved using the Americans for Democratic Action (ADA) scores. The objectives of our work are similar to those in Golbeck and Hansen (2014) but in this chapter we establish methods that work in a multi-party context, and, moreover, don't require a quantitative starting point, like the ADA scores or any other prior knowledge about the involved parties.

In Barberá (2015), the author uses the structural characteristics of the Twitter network to extract the political positions of politicians, users and news sources in five countries. He proposes a *Bayesian spatial following model of ideology* based on the popularity of the politicians, the political interest of users and their estimated ideal points on the political spectrum in order to predict the probability of a user following a politician. Although the author's hypothesis coincides with our hypothesis (i.e., the mere structure of the Twitter network suffices for the extraction of the political inclination of specific users), his proposed model applies extensive filtering on the users' dataset (e.g., geolocation, tweet activity, number of followers) while in our approach we use raw data for our algorithms without filtering and without any other knowledge of the users' characteristics.

Finally, in Ribeiro et al. (2018), the authors leverage the demographics of the audience of the news sources, obtained through the advertiser interfaces of social media sites like Facebook and Twitter, to infer biases of news sources. In a different work (Hannak et al., 2013), it is shown that opposing views of Twitter users can be reflected on the personalization of the corresponding Google News aggregator. In Le et al. (2017), a method for extracting information about the slant of a news article using related retweets and followers of Landmark users from Twitter, is presented. Selected Landmarks and connections and tweets from Twitter are used in An et al. (2012) with a global positioning algorithm to map news media on the political spectrum. The close association between MPs and news sources is also studied in Briola et al. (2018), where the political belief of a Twitter user is being inferred based on their links with news sources. All these related works are evidence that supports the view that there is significant political information in the Twitter network and that this information can be used to infer bias about news sources and, consequently, news articles. In this work, we show that political information can be extracted from Twitter even by using only the follower relations and that this information can be used to infer the bias and the affinity of news media.

3.3 Dataset and Methodology

In this section, we provide a description of the dataset and an overview of the methodology utilized to study the political affinity of the NOIs. Initially, a dataset is assembled from Twitter (Section 3.3.1), an online social network with distinctive political nature. We then explain how this dataset can be interpreted as a bipartite graph and suggest the appropriate transformation stage (projection) in order to reduce the dataset into

manageable size for the direct application of our algorithms (Section 3.3.2). Finally, we provide an overview of the proposed methodology in order to study the political affinity in the dataset (Section 3.3.3).

3.3.1 Dataset Description

The dataset that we assemble and use is based on the Twitter accounts of actors that are relevant to the Greek political scene. More specifically, we focus on (a) the members of the Greek Parliament (MPs), and (b) a list of the most acquainted news sources with nation-wide audience. We refer to these actors as NOIs (i.e., nodes of interest) since they occupy a significant share of information about the political scene in Greece.

The set of MPs was acquired from the official website of the Greek Parliament[1] without any discrimination. Summarily, the set of NOIs consists of 300 MPs, of which 166 have a public Twitter account that was either advertised in their personal websites or was a result of a query in the Twitter search engine. As a result, 134 MPs with either a protected account (5) or no account at all (129) could not be included in the dataset. Among the disregarded MPs is the party *KKE*, one of the eight political parties, representing the left wing of the Greek Parliament, of which none of the MPs have a Twitter account. Furthermore, 4 of the MPs were independent (not members of any party listing). We only considered the 162 MPs with an explicit party militancy as part of the NOI set (and not the 4 independent MPs). This decision was due to our methodology, which is based on a strict profile of political parties for both the evaluation of our experiments and the analysis of the news media affinity.

We also include a set of 24 well known news media in our dataset. In particular, the media contained in the dataset are:

1. 16 printed newspapers that are nationally distributed,
2. 6 TV channels with national broadcast range, and
3. 2 online blogs.

The selection of the news media is based on their nationwide coverage, their interest in political news, their presence in Twitter and on our commitment to cover, to the greatest possible extent, the political spectrum of Greece. We consulted a group of political scientists for advice on the coverage of the greek political scene by our dataset. We note that some well established news media of the Greek scene are not included in our dataset since they do not maintain, to the best of our knowledge, an official Twitter account. The political scientists confirmed that under these preconditions our dataset is representative of the political spectrum at that particular period. In total, we collected 186 Twitter accounts from the above categories. A breakdown of the NOIs is presented in Table 3.1.

[1] https://www.hellenicparliament.gr/en/Vouleftes/Ana-Koinovouleftiki-Omada/

Table 3.1: Breakdown of NOIs into groups. The MPs are shown in the left column and the news sources in the right. For the MPs the table shows the number of parliamentary seats for each party as well as the number of MPs present in the dataset. The underlined parties form the government coalition.

Parliamentary Group	Dataset	Seats	News Group	Dataset
SYRIZA	62	145	Newspapers	16
ND	61	76	TV Channels	6
DHSY	17	20	Blogs	2
XA	10	16		
KKE	0	15		
ANEL	4	9		
POTAMI	6	6		
EK	2	6		
Independent MPs	0	7		
Totals	162	300	Totals	24

We complete our collection by crawling the followers of each of the NOIs using the Twitter API to construct a dataset of 186 NOIs, 1,279,005 unique followers and 5,610,099 connections between NOIs and followers. For each of the users (NOIs and followers) only the Twitter user IDs are stored, while the connection is simply a pair of a NOI ID and a follower ID. It is worth mentioning that during this process we ignore the connections where both endpoints are NOIs. The rationale behind this decision is associated with our proposition to determine the political affinity of the NOIs without using information provided by their own actions directly. However, the amount of the NOI-to-NOI relations were less than 1% of the total following relations. Moreover, the 4 independent MPs as well as the additional news sources that are not included in the NOIs set are presented in the dataset as followers.

Finally, as a result of our data acquisition process, ties among the followers and non-NOI users are not included in the dataset. The process of obtaining this information is very demanding given the massive amount of followers and limitations imposed by the Twitter API but it is not considered necessary either since our methods do not leverage these connections.

The dataset was constructed on April 2018 and, thus, reflects the connections between the selected NOIs and their followers in the Twitter network, and consequently the political background, at that time. The dataset along with other supplementary material about this work are available online[2].

[2]https://doi.org/10.17632/jvfjdkhr5p

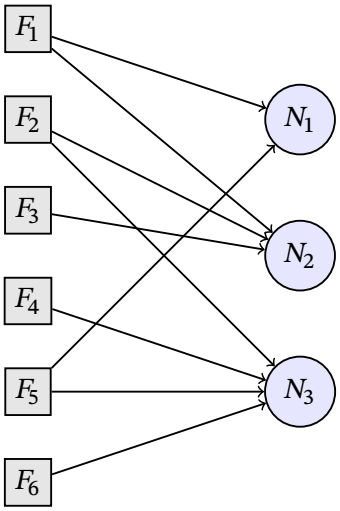


Figure 3.1: The projection of the bipartite graph onto the NOIs.

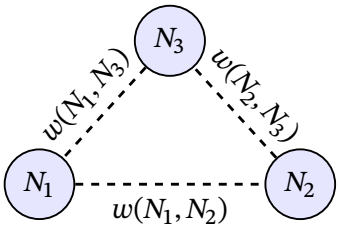


Figure 3.2: The projection of the bipartite graph of Figure 3.1 onto the NOIs. The weights of the projection edges are determined by a weighting method.

3.3.2 The Projected Graph

The acquired dataset can be naturally represented as a bipartite graph $\mathcal{G}(N, F, E)$ where the two disjoint sets of vertices are the NOIs (N) and the followers (F) respectively, and an edge E_{ij} between a NOI i and a follower j exists if and only if j is following i and j is not a NOI. Many real world networks are naturally modeled as bipartite graphs, especially in social systems, like the Twitter follower network we use in this work. An abstract example of a bipartite representation of the Twitter follower network can be seen in Figure 3.1. It is worth mentioning that the NOI with the greatest amount of connections in the graph is the user account of the –current by the time of the study– Prime Minister *atsipras* with 586,558 followers.

The dataset, however, is massive and possibly incompatible with general graph processing algorithms due to its bipartite nature. Thus, we transform the graph to its one-mode projection onto the NOIs, an extensively used method for compressing information about bipartite networks (Zhou et al., 2007). The one-mode projection of a bipartite network $\mathcal{G} = (X, Y, E)$ onto X (X projection for short) is a weighted, complete, unipartite network $\mathcal{G}' = (X, E')$ containing only the X nodes, where the weight of the edge between i and j is determined by a weighting function $\beta_{\mathcal{G}}(X_i, X_j)$. The weighting method may not necessarily be symmetrical but in this work we engage in a simpler approach with commutative weight functions so that $\beta_{\mathcal{G}}(X_i, X_j) = \beta_{\mathcal{G}}(X_j, X_i)$, resulting in an undirected projection. Typically, the weight function expresses a form of similarity among the vertices in order to preserve the semantics of the original graph. An example of a bipartite projection can be seen in Figure 3.2 as a complete graph consisting of the

Table 3.2: Notation used in the projection weighting methods in Section 3.3.2.

Set	Cardinality	Description
N	n	All the followers in the dataset.
N_z	n_z	The followers of NOI z.
N_{xy}	n_{xy}	The common followers of NOIs x and y, $N_{xy} = N_x \cap N_y$.

NOI vertices.

While the projection allows simplification of the network and compatibility with unipartite algorithms, it constitutes a lossy graph compression operation and consequently incurs information deficit over the original bipartite graph. There exists, however, no global weighting method of minimizing information loss and the optimal weighting is heavily dependant on the nature of the network and the objectives of the study. Therefore, we proceed with a selection of set theoretic functions that expose the similarity among the NOIs, namely the *Overlap coefficient*, the *Jaccard index*, the *Ochiai coefficient*, the *Sørensen-Dice coefficient* and the *phi coefficient*. These measures are briefly explained below. The notation that is used in the formulas is described in Table 3.2. Since all our weighting methods rely only on the follower sets, it holds that for any weighting method $\beta_{\mathcal{G}}$, if $N_x = N_y$, then $\beta_{\mathcal{G}}(x, z) = \beta_{\mathcal{G}}(y, z)$. This property is easy to prove via the following definitions.

Jaccard Index The Jaccard index of nodes x and y is defined as the intersection of the nodes' follower sets over their union:

$$j_{\mathcal{G}}(x, y) = \frac{|N_x \cap N_y|}{|N_x \cup N_y|}.$$

It has values in $[0, 1]$, with 0 signifying no common follower and 1 an equivalence in the follower sets.

Ochiai Coefficient The Ochiai coefficient between two NOIs x and y is identical to the cosine similarity when applied to binary vectors (presence or absence of an edge):

$$c_{\mathcal{G}}(x, y) = \frac{n_{xy}}{\sqrt{|N_x||N_y|}}.$$

The Ochiai coefficient can be described as the intersection over the geometric mean and is also a measure lying in $[0, 1]$.

Sørensen-Dice Coefficient The Sørensen-Dice coefficient is also known as the F1 score and is another statistic used for comparing the similarity of two follower sets:

$$s_{\mathcal{G}}(x, y) = \frac{2n_{xy}}{|N_x| + |N_y|}.$$

It can be shown that there is a relationship between Sørensen-Dice coefficient and the Jaccard index:

$$s_{\mathcal{G}}(x,y) = \frac{2j_{\mathcal{G}}(x,y)}{1 + j_{\mathcal{G}}(x,y)}.$$

As in the above methods, the Sørensen-Dice coefficient is in $[0,1]$ and is equal to the intersection over the arithmetic mean of the sets.

Phi Coefficient The phi coefficient is equivalent to the Pearson correlation coefficient when applied to binary vectors and is formulated as:

$$\phi_{\mathcal{G}}(x,y) = \frac{nn_{xy} - n_x n_y}{\sqrt{n_x n_y (n - n_x)(n - n_y)}}.$$

This measure differs from the other similarity functions as it can be in the range $[-1,1]$, where 1 is total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation. In some scenarios, however, a negative weight is either not meaningful or not compatible with the setting at all. In these cases we use two phi coefficient transformations instead that eliminate any negative value:

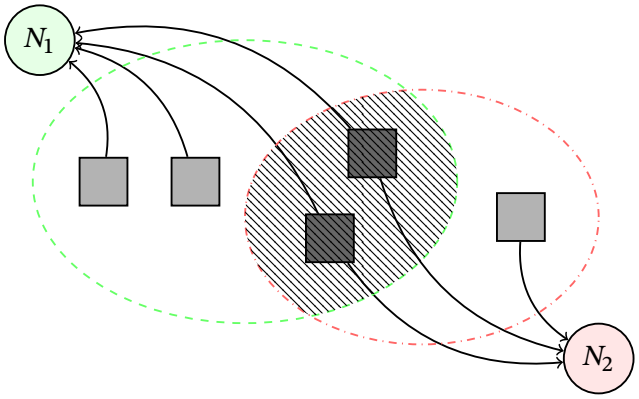
$$\phi_{\mathcal{G}}^{add}(x,y) = \phi_{\mathcal{G}}(x,y) + 1$$
$$\phi_{\mathcal{G}}^{exp}(x,y) = e^{\phi_{\mathcal{G}}(x,y)}.$$

Overlap Coefficient The overlap coefficient (also known as Simpson coefficient) is a measure in $[0,1]$ that measures the overlap between two sets. In our context it assumes the value of 1 if the nodes are identical and a value of 0 if they have no common follower. More specifically, it is defined as the size of the intersection divided by the smaller of the cardinalities of the two follower sets:

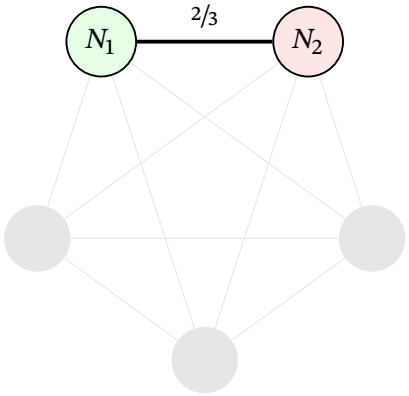
$$h_{\mathcal{G}}(x,y) = \frac{|N_x \cap N_y|}{min(|N_x|, |N_y|)}.$$

A visual example of the application of the overlap coefficient is shown in Figure 3.3. In that example, the follower connections of NOIs N_1 and N_2 are being utilized to compute their overlap coefficient, which is 2/3.

All weighting methods express similarity and are, hence, generally positively correlated. However, each projection method interprets the concept of similarity from a different perspective and, in that sense, they are all unique. Other similarity techniques mentioned in the literature are computationally demanding, for example the original SimRank (Jeh & Widom, 2002) algorithm has a space requirement of $\mathcal{O}(n^2)$. In this work, we utilize methods that can be computed efficiently even for large scale input data, such as the Twitter network. Our observations suggest little room for further improvement over the effectiveness of these simple similarity measures.



(a) Illustration of the followers of N_1 and N_2 as a Venn diagram.



(b) The edge between N_1 and N_2 in the projection is equal to the overlap coefficient between them (2/3).

Figure 3.3: Abstract example of the application of the overlap coefficient for two NOIs N_1 and N_2 to create a weighted projection of the bipartite graph.

The projection methods were applied on the bipartite graph G so that the weight of the projected edge between two NOIs x and y is evaluated by a function $weight(E'_{xy}) = \beta_G(x, y)$. We did not take self loops into consideration because the weights would be trivially set to the maximum value. Since the projected graph is complete, there are $|N|(|N|-1)/2 = 17{,}205$ undirected edges in each of the projections, including possible edges with zero weight.

3.3.3 Methodology

Our proposed methodology utilizes the projections of the Twitter follower network in order to estimate the political affinity of the NOIs. Specifically, our methodology includes a combination of methods, namely the *modularity clustering*, the *minimum linear arrangement (MinLA) problem* and the *DeGroot model approach with stubborn agents*. The selection of these methods highlights the conceptual diversity in the interpretation of political affinity. Each of these approaches provides a different perspective of the political affinity encoded in the projection graphs.

Modularity Clustering

Clustering or *community detection* in a graph refers to the process of identifying the modules and, possibly, their hierarchical organization, by using only the information encoded in the graph topology. Community detection has been widely applied in real-world social systems and various methods with different characteristics have been suggested (Fortunato, 2010). More information about community detection in general has been given in Section 2.5.2.

More specifically, we use the algorithm in Blondel et al. (2008), a heuristic method that

is based on modularity optimization and is commonly known as *Louvain optimization*. The algorithm is well established and has seen extensive use in the field of social networks (Lancichinetti & Fortunato, 2009). Its complexity is linear on typical and sparse data. The Louvain optimization algorithm unveils hierarchies of communities and allows to zoom in the network and to observe its structure with the desired resolution via the parameter r. Therefore, the *resolution* parameter determines the desired number of communities in the partition. The parameter can be tuned accordingly in order to accommodate the requirements of specific experiment settings.

The application of modularity clustering to the dataset enables us to study the political affinity of the NOIs from a perspective that conveys the group affiliation, i.e. the affiliation of a NOI with a specific political party. Thus, our expectation is the partition of the NOIs into sets of communities that represent the political parties.

Minimum Linear Arrangement

The *Minimum Linear Arrangement* (MinLA) problem consists in finding an ordering of the nodes of a weighted graph, such that the sum of the weights of its edges is minimized. More formally, given a finite graph $\mathcal{G} = (V, E)$ of order n with weighted adjacency matrix w, the MinLA problem is the problem of finding a vertex labeling $f \to \{1, 2, ..., n\}$ such that the sum

$$\sum_{(u,v)\in E} w_{uv}|f(u) - f(v)|$$

is minimized over all possible labelings (Safro et al., 2006). More information about the MinLA problem has been given in Section 2.5.3.

The MinLA problem has been applied to various scientific fields, for example in VLSI design (Petit, 2003) in order to minimize the electrical resistance of a circuit, or in a theoretical level (Chierichetti et al., 2009) but, to the best of our knowledge, its physical interpretation has not been studied on a specific social network theme in prior literature. We argue that the MinLA problem is suitable for application on this context and constitutes an innovative approach for the analysis and understanding of social networks. Our hypothesis is that the application of a solution of the MinLA problem to the graph projections will unveil the positioning of the NOIs in a bipolar spectrum and, eventually, in a political spectrum. The intuition behind this is that NOIs with similar political views and, hence, stronger bonds in the projection, should occupy successive labels in the MinLA ordering, while NOIs that share weaker links should be distantly positioned.

For the purposes of this work, we designed and implemented a randomized local search algorithm, repeated over a set of uniformly random initial rankings, which approximately leads to the minimum cost linear arrangement (LA). Given an initial guess of the arrangement we perform a sequential series of steps to determine a local

minimum of the cost function, the *fast converge phase* and the *local converge phase*. During the fast phase, the algorithm performs random pairings on the elements of the arrangement for a number of repetitions, and swaps the elements of a pair if it improves the cost. The purpose of this phase is to allow the algorithm to quickly descend close to a local minimum while the number of repetitions involved determine the convergence rate. We selected to perform this step n^2 times as we empirically observed a sufficiently quick convergence for this setting. During the local phase we validate that the current LA is the local minimal cost LA by performing all possible swaps in it; if there is a swap that improves the cost we restart the process until we identify the local minimum. The above process of computing a local minimum is repeated several times with random initial arrangements and the best solution is kept. The scheme is presented in Algorithm 1.

Algorithm 1 Local Search MinLA algorithm

```
1:  procedure LOCALMIN(a: Array)
2:      Set n ← size(a)
3:      for n^2 times do                                  ▷ Fast converge
4:          Perform a random swap on a to create a'
5:          If it reduces the cost set a ← a'
6:      while changed do                                  ▷ Local converge
7:          Set changed ← false
8:          for x in [1, n), y in (x, n] do
9:              Perform the swap (x, y) on a to create a'
10:             If it reduces the cost set a ← a' and changed ← true
11:
12: procedure MAIN(a: Array, reps: Int)
13:     for reps times do
14:         Shuffle a to create a' and invoke LOCALMIN(a')
15:         If the cost of a' is lower than a set a ← a'
```

The DeGroot Model Approach

The DeGroot model is an opinion diffusion model introduced by Moris H. DeGroot (Degroot, 1974), whose core idea is that individuals tend to adopt the opinions of their friends. We have discussed the DeGroot model in more detail in Section 2.5.4, where we also showed the details regarding the analysis by Ghaderi and Srikant (Ghaderi & Srikant, 2013) who enriched the DeGroot model with stubborn vertices.

Based on the findings of Ghaderi and Srikant, we present a technique to estimate the political affinity of the NOIs. We consider each NOI projection as a social graph where the weights of the links produced by the projection methods correspond to the trust factors of the nodes to their neighbors. In the case of the phi projection,

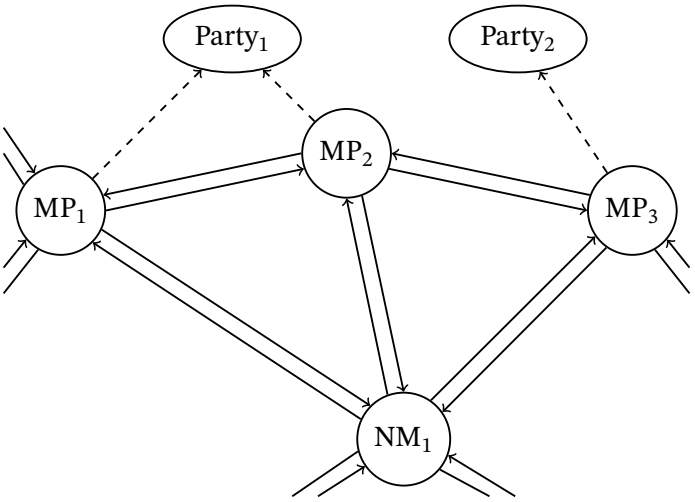


Figure 3.4: Abstract example of a NOI projection, enriched with party nodes.

the graph contains edges with negative weights, a phenomenon that does not abide by the restrictions of the DeGroot model. Therefore, transformations of the original formula are utilized (namely the $\phi_{\mathcal{G}}^{add}$ referred as Phi-A and $\phi_{\mathcal{G}}^{exp}$ as Phi-E) as described in Section 3.3.2. The undirected edges of the graph are duplicated into two opposite directed arcs and, in order to abide by the DeGroot model restrictions, the weights of each node's outgoing arcs are normalized. Seven additional nodes are introduced to the graph that represent the political parties and each MP's node is linked to the corresponding party.

Figure 3.4 presents an abstract example of a NOI projection with three MP nodes (MP$_1$, MP$_2$, MP$_3$) and one news media node (NM$_1$), enriched with two party nodes (Party$_1$, Party$_2$). According to the needs of individual experiments, we perform slight modifications to this structure to ensure compliance with the respective evaluation goals. Specifically, depending on the setting, a MP can be transformed into a stubborn node by removing its outgoing arcs to other NOIs and have its sole influence exerted by its respective party. Conversely, when a MP updates its opinion according to the DeGroot model, its friendships to other MPs (outgoing compact arcs) are used and its link with the party is ignored. Hence, the direct link of a MP to its party and the MP's friendships are mutually exclusive.

The use of random walks leverages the connections among the MPs as well as the links between the MPs and their parties to quantify the latent relationships of MPs with all political parties. Consequently, our heuristic can also uncover the associations among the news media and the parties through the intermediate edges with neighbouring MPs in the projections. This method reveals a perspective of political affinity that differs from the NOI clustering and the MinLA approach as it relies on the hitting probabilities of random walks to determine the influence factors between the various actors in the political graph.

3.3.4 Alternative Dataset Usage

The dataset and the projections described in this section were also utilized in joint works during the research of this thesis. The rich political information within the dataset and the simplification of the network with the application of projections establishes opportunities to study the political affinity of nodes using different perspectives.

Initially, in Briola et al. (2018), it is demonstrated that the political beliefs of a user in the Twitter network might be possible to be revealed using the connections of that user, i.e. their friends and followers. The privacy leakage is determined by the accurate prediction of the MPs that the Twitter user is following or is predicted to be linked to. For this reason, two approaches are developed, where the core idea is that some of the connections of a Twitter user might be able to predict the other connections, which is inline with the homophily property. In the first approach, the used method is able to predict the MPs followed by the Twitter user in question by utilizing that user's followers, where this user is not necessarily a high profile person of political interest. In the second approach, the MPs that are followed by a Twitter user are being predicted based on the news media followed by that user. The second approach demonstrates the property of *selective exposure* and the consistency of users' actions with respect to their beliefs.

The core study of Gyftopoulos et al. (2020) is the political impartiality of news media in multipartite political scenes. The *political impartiality* of a news outlet is determined by the lack of political bias, i.e. a completely impartial news source is the one whose political bias cannot be discerned by the content of their stories. The methodology developed in this work aims to quantify the political impartiality of news media using the Twitter dataset (Greek MPs and popular news media) and its projections that were described previously in this section. The popular news sources in the dataset are ranked according to their deviation from the ideal impartial medium and the results are evaluated via an online survey given to a group of political scientists.

The application and experimentation with the DeGroot model approach presented in this chapter are in collaboration with Sotirios Gyftopoulos. The results are presented in this thesis for consistency and comparison with the rest of the methodology.

3.4 Proof of Concept: The MPs Case

The fundamental concept of our study is that the mere structure of a social network consisting of nodes with political interests suffices for the extraction of rich political information through the use of innovative algorithms. In order to confirm this assertion, we firstly apply our proposed methodology on a subset of the acquired dataset that ensures, to the greatest possible extent, the existence of political information and confines any source of politically irrelevant information that may falsify the results

of our methods. We consider this as the proof of concept scenario of our proposed methodology.

3.4.1 The MPs Dataset and Projections

In the MPs case, we use a subset of the acquired dataset that contains the MPs and their Twitter followers as well as the connections between them. In particular, the dataset includes 162 NOIs (i.e., the Twitter accounts of the MPs), 740,580 followers and 2,403,200 connections between NOIs and followers. We note that the news media presented in Section 3.3.1 are considered as mere followers in the context of this scenario and any connection with the MPs is included. We can safely argue that this confined dataset incorporates to the greatest possible extent the available political information about the greek scene since the particular NOIs (i.e., the greek MPs) exhibit profound political behaviour and it is only natural to assume that their followers are politically motivated.

The raw data are perceived as a bipartite graph and are transformed into projected graphs using the projection methods described in Section 3.3.2. We refer to the bipartite graph and its projections as *MPs graph* and *MPs projections* respectively.

3.4.2 Experiments and Results

The proof of concept experiments utilize our methodology in order to confirm the existence of rich political information in the dataset. The application of *Modularity Clustering*, the *Minimum Linear Arrangement (MinLA)* and the *DeGroot Model Approach* are described separately and our results are being presented in the following sections.

Modularity Clustering

Initially, we apply the modularity maximization algorithm (Section 3.3.3) on the MP projection graphs in order to partition the vertex set into disjoint groups of MPs and show that this method can reveal the underlying political structure of our dataset. Our hypothesis is that modularity clustering will partition the MPs into their respective political parties, or, equivalently, that the MPs of the same party will be classified into the same cluster. Consequently, the evaluation of the clustering method is performed towards the true partition of MPs in political parties (Table 3.1) which is an objective indication about their political affinity. As a result, we tune the resolution parameter so that the algorithm returns 7 clusters, the amount of political parties in the ground truth.

The quantifiable evaluation can be achieved by reducing the partition correlation problem into a set similarity problem using the concept mentioned in Alzahrani and Horadam (2016, Section 2.2.1). More specifically, for some partition of the nodes $[N]$ into

Table 3.3: Clustering evaluation of the MP projections. Each projection displays the maximum (odd line) and minimum (even line) value of the respective evaluation measure.

$\beta_{\mathcal{G}}$	Evaluation measure					
	Jaccard	SMC	F1	NMI	Pearson	Cosine
Overlap	**.8277**	**.9456**	**.9057**	**.6671**	**.8693**	**.9066**
	.7867	.9314	.8806	.6040	.8346	.8816
Ochiai	.5449	.8436	.7054	.3148	.6117	.7118
	.4724	.8197	.6417	.2515	.5459	.6544
Phi	.5110	.8384	.6764	.3055	.5968	.6911
	.4358	.8133	.6071	.2427	.5276	.6298
Jaccard	.4763	.8244	.6453	.2663	.5584	.6608
	.3759	.7831	.5464	.1661	.4452	.5678
Sørensen	.4576	.8144	.6279	.2386	.5312	.6418
	.3981	.7901	.5695	.1810	.4664	.5880
Random	.1062	.6475	.1920	.0000	.0000	.2046

groups we consider the set S to comprise all unordered node pairs $\{i, j\}$, with $i \neq j$, where elements i and j belong to the same group in that partition and S' to consist of all other pairs. Naturally, it has to hold that $|S| + |S'| = |N| \cdot (|N| - 1)/2$.

The resulting sets S and S' can then be used as input to our evaluation methods, which are the Jaccard index, the Simple Matching Coefficient (SMC), the F1 score, the Normalized Mutual Information (NMI), the Pearson correlation and the Cosine similarity. These measures are used to assess the effectiveness of a partition and are different from the measures used to construct the projection, although some of them are used for both purposes. The results are shown in Table 3.3. The rows of the table refer to the similarity functions used for the projection. Each function is represented by the minimum and maximum values of the respective evaluation measure over all resolutions between 0.2 and 3.0 with a step of 10^{-3} that yielded 7 clusters. The columns of the table denote the evaluation measures. For comparison, the random partitioning is also included in the table. This random evaluation was produced separately for each measure/column by gradually generating random partitions of 7 communities (as many as the political parties) until the average of the correlations did not change beyond the 9th decimal digit.

The results deliver a strong evidence about the validity of our hypothesis, stating that MPs of the same political party will be classified into the same group in the partition. The strength of the correlations among the weighting methods varies, although all methods had an above random association. In particular, the overlap coefficient firmly outperforms other functions commonly mentioned in the literature on all evaluation measures. Therefore, this indicates the existence of rich political affinity information within the Twitter follower network, and substantiates the suitability of modularity clustering for obtaining this information.

Figure 3.5: Force-directed visualization of the MP overlap projection.

Figure 3.5 displays a force-directed visualization, produced by Gephi (Bastian et al., 2009), of the MP projection using the overlap function with a resolution of 0.855. These settings correspond to the highest correlation achieved (the first line of Table 3.3). Vertices in this layout are colored by their modularity class. The respective party of each node (the ground truth) is given as a text label within the node while the party IDs are given in the legend in the top left corner. The partition of the clustering illustrated in this figure is a result of only the Twitter follower-followee relations while the real distribution of MPs in political parties (the ground truth) is only used for the evaluation.

An important visual observation is that the accuracy of the identified clusters is remarkably high, which coincides with the results in Table 3.3. In particular, the biggest political parties (SYRIZA, ND, XA) are clearly identified with the respective clusters almost flawlessly. A further observation is that the nodes of A. Tsipras and K. Mitsotakis, the leaders of the two largest parties which are correctly classified, are located near the center of the visualization. A possible explanation is that nodes with large degrees also have a large portion of their followers choose them not because of their political identity, but simply because they are the leaders of the two largest parties. Additionally, the layout provides a visual perception of the close association between modularity clustering and force-directed placement (Noack, 2009).

Minimum Linear Arrangement

In this experiment we apply our MinLA algorithm to the MP projections in order to arrange the MPs in a one-dimensional space and study the significance of this ordering. Our hypothesis is that MPs of the same political party will appear consecutively inside the minimum cost arrangement of the MP projection vertices; the known affiliations of MPs in political parties enables us to evaluate this. Finally, we make an attempt to attribute the physical meaning of the minimum cost arrangement in relation to the left-to-right political axis.

Initially, we apply this algorithm to all of the MP projections and, for each, we obtained the minimum cost arrangement m and its cost $C(m)$. Since in this experiment we deal with ordinal data, we also define the concept of party ordering. Our dataset contains 7 parties, so there are $\rho = 7! = 5{,}040$ possible orderings. Each of these orderings can be flattened to a ranked list of MPs, where MPs of the same party are tied on the same rank. Thus, there are also ρ flattened MP ranked lists denoted as R_i, $1 \leq i \leq \rho$.

We assess the correlation of m with every ordering R_i using the Kendall tau-b (τ_B) correlation coefficient (Agresti, 2010), which is a statistic used to measure the ordinal association between two measured quantities. The tau-b correlation coefficient is a generalization of the Kendall tau-a coefficient that accounts for ties in the input lists, specifically present in the R_i orderings. It is worth noting that Kendall tau-b is in range $[-1, 1]$ but, since in our context the linear arrangements (LAs) cannot contain ties, the

$\beta_{\mathcal{G}}$	Min. (Found) Cost	Random Cost	τ_B
Overlap	162,196	225,083	0.7261
Phi	51,230	83,352	0.7228
Ochiai	55,308	88,322	0.7008
Jaccard	19,628	38,216	0.3933
Sørensen	36,624	68,293	0.3867

Table 3.4: MinLA evaluation of the various MP projections. The max τ_B is 0.8361 while the random τ_B is approximately 0.1149. The costs among the projections refer to different edge weights and, thus, are not comparable.

maximum value is

$$K_{max} = \frac{T(n) - \sum_i T(t_i)}{\sqrt{T(n)}\sqrt{T(n) - \sum_i T(t_i)}}, \text{ where } T(x) = \frac{x(x-1)}{2},$$

which equals 0.8361 because $t = [62, 61, 17, 10, 6, 4, 2]$ (Table 3.1). Afterwards, we find the party ordering with the highest correlation to m, defined as R_q where

$$q = \arg\max_{i \in [1,\rho]} \tau_B(m, R_i).$$

Our results are presented in Table 3.4, which displays the tau-b correlation of each projection's minimum cost LA against its respective R_q. The results are in agreement with our findings in Section 3.4.2 in regards to the effectiveness of the overlap projection and the existence of rich political information within the Twitter follower dataset. Specifically, the $\tau_B(m, R_q)$ of the overlap coefficient is 86.8% of the maximum (0.7261/0.8361) proving that the MinLA problem definition highlights the clustering features of our dataset and confirms our hypothesis. Moreover, the phi and Ochiai based projections are also represented by very promising correlations that are only marginally below overlap. The random cost column of Table 3.4 is derived from the average distance of two nodes in a random LA which is $(n+1)/3$ and is stated in the following lemma.

Lemma 1. *The average distance of two nodes in a random LA is* $(n+1)/3$.

Proof. Let X and Y be two random variables for the positions of the two nodes, respectively, in the LA. First, assume $X < Y$. Then, the following sum S_1 is:

$$\sum_{x=1}^{n} \sum_{y=x+1}^{n} P[X = x] \cdot P[Y = y | X = x] \cdot (y - x)$$

$$= \frac{1}{n} \frac{1}{n-1} \sum_{x=1}^{n} \sum_{y=x+1}^{n} (y - x) = \frac{1}{2n(n-1)} \sum_{x=1}^{n} (n - x)(n - x + 1)$$

Assuming $X > Y$, the corresponding sum S_2 has the same value $S_2 = S_1$. The average distance is equal to the sum $S_1 + S_2$. Adding S_1 and S_2, and then simplifying gives $(n+1)/3$. □

(a) Resulting minimum cost LA.

(b) Closest party arrangement.

Figure 3.6: Visualization of the minimum LA for the overlap projection. The top figure is the resulting minimum cost LA and the bottom one of the closest party arrangement for it. The τ_B correlation between the two arrangements is 0.7261. Due to its nature this figure might not be readable in grayscale form.

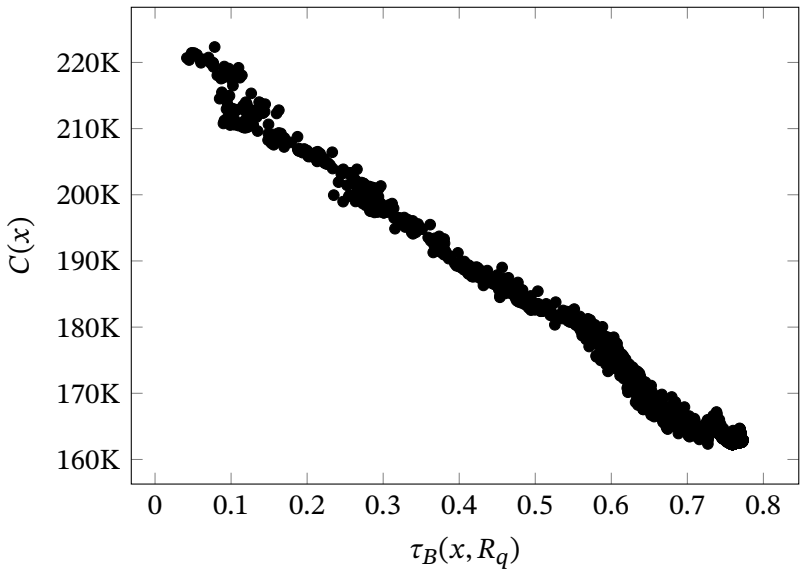


Figure 3.7: Convergence of the arrangement cost as a function of the correlation with R_q.

Furthermore, we experimentally found that the $\tau_B(r, R_q)$ of the random MP ordering r is 0.1149. Consequently, it follows that, despite their differences, all of the projection methods reveal some amount of information from the follower network with above random significance.

A visual perception of the MinLA application of the overlap projection is shown in Figure 3.6. The top ruler in the figure displays the minimum cost LA m while the bottom represents the closest party arrangement R_q. Each ruler contains 162 MPs represented by points colored by the real party of the respective MP. The figure offers an alternative understanding of the magnitude of correlation between these vectors and, overall, the validity of our hypothesis.

Another observation stems from Figure 3.7, which shows the relation between the MinLA cost $C(x)$ and the tau-b correlation $\tau_B(x, R_q)$ of the current minimum cost LA x of the overlap projection as the algorithm converges. The south-most point in this figure is m. This figure also shows that there is a very strong, almost linear relationship of the LA cost function with the correlation function.

Finally, we discuss some interesting observations about the closest political party or-

dering R_q of the overlap projection which is [SYRIZA, ANEL, POTAMI, DHSY, ND, EK, XA]. A comparison of the R_q to the arrangement of the parties based on their ideological identity reveals interesting properties of our result and of the inherent political information in our dataset. According to their self-identification and data from additional resources (e.g. Wikipedia), the most credible arrangement of the parties on the left-to-right political spectrum is [SYRIZA, DHSY, POTAMI, EK, ND, ANEL, XA]. In R_q, ANEL is adjacently positioned to SYRIZA, an oxymoron phenomenon that can be justified by the fact that ANEL and SYRIZA were in governmental coalition and, thus, their ties are strong in the Twitter follower dataset. Furthermore, the swap of DHSY and POTAMI in R_q is inconsequential, especially after their deliberations (in April 2018) about the formation of a new upcoming coalitional party (KINAL) for the upcoming elections. The misplacement of EK can be attributed to its small footprint (2 MPs) and, hence, by deficient information. In general, we can argue that R_q outlines the parties on one dimension according to the followers' criteria that are a combination of the left-to-right political perspective and the pro and anti-government feeling.

The DeGroot Model Approach

In this section, our DeGroot model approach is applied to the MP projections in order to determine the influence factors towards the political parties. These factors are then used to classify every MP to the party with the maximum influence factor. Our hypothesis coincides with the clustering hypothesis in Section 3.4.2; MPs will have their dominant influence factors on their respective affiliated party.

We perform a series of experiments for all the MP projections that are based on the concept of the leave-one-out cross-validation method, where each MP is selected individually. The directed arc of the selected MP to its party is ignored while the rest of MPs are transformed into stubborn agents by removing their outgoing arcs to other MPs as explained in Section 3.3.3. The selected MP's friendships with other NOIs are used to calculate a random walk's hitting probabilities to every party's node given it originates by the MP. Since the parliamentary groups are uneven, the evaluated probabilities are divided by the corresponding group's size in order to compute the *uniform per party influence* and avoid any dominance effect by the parties with large parliamentary groups. The uniform influences are used to classify each MP to a party based on the greatest uniform influence of their random walk. A hit is considered when the party with the greatest uniform influence on the MP coincides with its actual party. The experiments are implemented using PRISM (Kwiatkowska et al., 2011), a tool that is widely used to analyze models that exhibit probabilistic behavior (e.g. Markov chains).

The results presented in Table 3.5 denote that our approach achieves surprisingly high hit ratio in almost all cases of projections, a clear indication that the MPs dataset and, consequently, the MPs projections contain significant political information. The highest hit ratio is achieved in the graph produced by the Ochiai projection (88.89%) while the

Table 3.5: Hits of the leave-one-out cross-validation method for each projection.

	Overlap	Ochiai	Jaccard	Phi-A	Phi-E	Sørensen
SYRIZA	53	52	48	49	49	46
ND	59	57	53	55	57	53
DHSY	16	16	13	7	9	13
XA	10	10	10	10	10	10
ANEL	3	3	0	0	0	0
POTAMI	1	6	6	0	0	0
EK	0	0	0	0	0	0
Total Hits	142	144	130	121	125	109
Hit ratio	87.65%	88.89%	80.24%	74.69%	77.16%	67.28%

result of the corresponding overlap projection is slightly lower (87.65%). Moreover, the Ochiai and the overlap projections provide sufficiently robust graphs that enable the correct classification of MPs of smaller parties (e.g. ANEL, POTAMI). This property is also valid for the graph of the Jaccard projection although it exhibits noticeable weaknesses in the classification of MPs of the two largest parliamentary groups (SYRIZA and ND). In general, the vast majority of the MPs projections produce graphs that achieve high levels of accuracy in the classification of MPs to their parties.

3.4.3 Discussion

The results of all three methods presented in the previous sections provide clear indications that our assembled dataset contains significant political information and the applied algorithms are efficient in extracting it. The selected methods succeeded in revealing different aspects of the political information. The results of modularity clustering indicate that the followers of NOIs suffice for the efficient detection of the actual parties while the minimum linear arrangement produces a ranking of the NOIs that can be interpreted as a political bipolar. Finally, the DeGroot Model Approach exhibited surprisingly successful behavior in highlighting the affiliations of the NOIs to the political parties.

The results of our proposed methods also allow a comparative analysis of the weighting projection methods and their efficiency in conveying useful information in the projections. The phi coefficient provided efficient scores in all methods applied and the Ochiai weighting method achieved the highest scores in the DeGroot Model Approach. In general, the overlap coefficient appears superior to other measures. Although, it does not always attain the best evaluation in all the scenarios, it qualifies for a very consistent and reliable weighting function among the proof of concept experiments in our dataset.

3.5 The Case of News Media

The promising results of our methodology on the purely parliamentary dataset of the previous section justify our attempt to deploy the presented algorithms on a politically obscure scene. We consider a set of popular news media in Greece as the case study and utilize our proposed methodology to examine its effectiveness. For evaluation purposes, we conduct an experts' survey and compare its findings to the results of our algorithms.

3.5.1 The News Media Dataset and Projections

In this case study, we utilize the complete dataset of our work. The dataset includes 186 NOIs (MPs and news sources), their followers and the connections between NOIs and followers (see Section 3.3 for further details). The bipartite graph formed by this data is projected onto the NOIs using the projection methods of Section 3.3.2 to create the *enriched projections*, which are then provided as input to the methods presented above.

3.5.2 Expert Survey

The purpose of the expert survey is to establish a factuality that we consider as ground truth about the political affinity of the news media that are present in our case study. We resort to the expertise of 8 scientists from the field of political science to provide us insight about the political affinity and orientation of news sources in Greece. We structured and provided a survey questionnaire about the 24 news media of our dataset.

The questionnaire involved two questions, which we refer to as *political affiliation* and *political orientation*. The first question aimed at providing means of quantifying the relationship among news sources and political parties. The participants were asked to label the relationships among all pairs of news sources and political parties with one of the options "-2 hostile", "-1 negative", "0 neutral", "1 positive" and "2 partisan". The second question aimed at classifying the news sources in a left-to-right political spectrum scale. The experts were asked to label each news source with one of the options "far left", "left", "center", "right" and "far right". These options represent the position of the news media in the political spectrum and do not directly imply an association with a political party as in the first question.

The responses of the participants are aggregated using the average of individual answers. The data of the first question were processed into a 24×8 matrix (24 news media and 8 political parties) of political affiliation values in the range [-2,2] while the responses of the second question were assigned values in the range [-2,2] (i.e., -2 = "far left", -1 = "left", 0 = "center", 1 = "right", 2 = "far right") and the average values denote the political orientation of each news media derived from the opinions of the experts.

We consider these findings of the survey questionnaire as the ground truth that we can deploy in the application of our methods in the news media case study. The raw answers of the survey are included in the supplementary material of this work given in Section 3.3.1.

The reliability and homogeneity of the expert survey was assessed using Cronbach's alpha (Krippendorff, 2018). The issue of missing data (245 of 1728 records) was solved using a variety of imputation techniques (Pigott, 2001) (namely the k-nearest neighbours, multivariate imputation by chained equations (MICE), expectation maximization (EM), mean, mode, median and random imputation) and listwise deletion. The Cronbach's alpha values that were calculated for all these missing data handling methods ranged from 0.929 to 0.956 indicating acceptable (≥ 0.7) internal consistency of the conducted survey.

3.5.3 Experiments and Results

The final step of our case study involves the application of the Modularity Clustering, the Minimum Linear Arrangement (MinLA) and the DeGroot Model Approach to the enriched projections. The outputs of the algorithms outline the political profile of the news media under different prisms. The results are evaluated using the findings of the experts' survey.

Modularity Clustering

The methodology explained in Section 3.4.2 is reproduced for the 5 enriched projections and from the resulting partitions the news sources are filtered. It is then possible to use the expert responses of the political affiliation question for the evaluation of the clustering method. Specifically, we assign each news source to a cluster based on the political party that is most affiliated with that source and, hence, creating a comparable structure.

This process yields 5 groups of news sources but one of the news sources is tied in two parties, one party with 13 NOIs (including the tied news source) and a singleton community comprising only the tied news source. However, since the evaluation method relies on pairs of nodes inside the clusters, it is not able to distinguish a singleton cluster. Furthermore, the method cannot operate on overlapping partitions and, thus, we naturally place the tied news source into the bigger community, eliminating the singleton group. This action is inconsequential since that particular node has the same level of political affiliation for both parties. Therefore, the resolution parameter of the clustering method is, similar to Section 3.4.2, tuned for 4 communities.

The results are shown in Table 3.6, which has the same format as Table 3.3. For each projection method, the minimum and maximum values of the respective evaluation measure over all resolutions that yielded 4 clusters is displayed. For comparison, the

$\beta_{\mathcal{G}}$	Evaluation measure					
	Jaccard	SMC	F1	NMI	Pearson	Cosine
Overlap	**.5072**	**.7536**	**.6731**	**.1736**	**.4762**	.6734
	.4488	.7355	.6196	.1467	.4367	.6226
Ochiai	.4747	.6884	.6438	.1406	.3927	.6768
	.2865	.4638	.4455	.0009	.0343	.4470
Phi	.5029	.7065	.6692	.1566	.4386	**.6865**
	.2825	.5000	.4405	.0026	.0588	.4432
Jaccard	.4067	.6957	.5782	.0810	.3241	.5787
	.3033	.4674	.4655	.0000	.0051	.4698
Sørensen	.4067	.6812	.5782	.0764	.3187	.5832
	.2757	.4891	.4322	.0004	.0240	.4368
Random	.1722	.5685	.2927	.0000	.0000	.2987

Table 3.6: Clustering evaluation of the enriched projections. Each projection displays the maximum (odd line) and minimum (even line) value of the respective evaluation measure.

random partition with 4 communities is also given. The table carries similarities with the experiments on the MP projections. More specifically, given the significance of the measures, the existence of political information within the Twitter follower dataset is further established. It also appears that the modularity optimization clustering is suitable for the examination of the political affinity of the news sources within our dataset. Moreover, the overlap similarity appears to retain more information about the objectives of this experiment, although all of the projections achieved better than average accuracy.

Minimum Linear Arrangement

In Section 3.4.2, the application of the MinLA problem in the MP dataset (through the MP projections) demonstrated the suitability of our methodology as well as the existence of profound political information within the Twitter follower network. For the purposes of the case study, we apply the same methodology on the projections of the enriched graph. Our motive is to confirm the previous findings in Section 3.4.2 and to examine new hypotheses about the news sources by utilizing the results of the expert survey.

More specifically, the application of the same MinLA algorithm in each enriched projection yields an arrangement m of the NOIs in a line, which is similar to the arrangement in Section 3.4.2 but in this case study it contains the news sources in addition to the MPs. It is possible to use m as source to construct two sub-arrangements m_{mps} and m_{news} which consist of only the MPs and news sources respectively, where the relative ordering of the NOIs is preserved inside the sub-arrangements. In the undermentioned text, we study these two sub-arrangements separately.

Initially, we evaluate m_{mps} against the MP distribution in the political parties in the

Table 3.7: MinLA evaluation of MPs using two datasets.

$\beta_{\mathcal{G}}$	τ_B of MP projections	τ_B of enriched projections
Overlap	0.7261	0.7587
Phi	0.7228	0.7617
Ochiai	0.7008	0.4315
Sørensen	0.3867	0.3809
Jaccard	0.3933	0.3875

Table 3.8: MinLA evaluation of the various enriched projections. The max τ_B is 0.9799. The costs among the projections refer to different edge weights and, thus, are not comparable.

$\beta_{\mathcal{G}}$	Minimum (Found) Cost	Random Cost	τ_B
Overlap	260,787	351,758	0.6249
Phi	71,732	115,967	0.4030
Ochiai	77,727	124,053	0.4030
Sørensen	46,512	93,088	0.1960
Jaccard	24,816	51,930	0.1738

same way as in Section 3.4.2 while also contrasting the results. The two arrangements differ only on the dataset used and, thus, this evaluation shows how the addition of the news sources affected the political information in the dataset. The results are shown side by side in Table 3.7. It is clear that the addition of the news sources in the dataset did not diminish the amount of political information within the dataset. In fact, the projection methods that did well with the MPs dataset (overlap and phi), also performed well in the enriched dataset. This is a strong indication that the clustering features within the Twitter follower network are maintained even after the enrichment with the news sources. Furthermore, the Sørensen and the Jaccard projections had negligible differences while the Ochiai projection received a significant drop in effectiveness.

The evaluation of m_{news} is performed against the replies about the political orientation in the expert survey. Specifically, the political orientation vector given in Section 3.5.2 indirectly creates a ranked list of the news media sorted by their orientation. The evaluation of m_{news} is performed against this ranked list. This process is semantically very different from the m_{mps} evaluation. More precisely, the political orientation of the news sources corresponds to a linear scale of the news sources in the left-to-right political spectrum. As such, the evaluation of m_{news} is an indication of the minimum cost MinLA news sources arrangement correlation with the left-to-right political spectrum orientation.

The results of our experiment are reported in Table 3.8, which has the same form as Table 3.4. The max τ_B was calculated using the formula given in Section 3.4.2; due to the low amount of ties, the max τ_B is much higher than the respective value in Section 3.4.2. The table confirms some of our previous findings regarding the effectiveness of the overlap projection and further demonstrates the potency of our methods.

	Purely parliamentary graph	Enriched graph
Overlap	87.65%	88.27%
Ochiai	88.89%	87.04%
Jaccard	80.25%	80.25%
Phi-A	74.69%	72.84%
Phi-E	77.16%	74.69%
Sorensen	79.01%	77.78%

Table 3.9: Hit ratios of MPs based on the leave-one-out cross-validation method for purely parliamentary (only MPs) and enriched projections (MPs and news sources).

The inspection of the τ_B correlation between the news sources minimum cost arrangement and the expert survey political arrangement provides interesting insights. More specifically, although the result is clearly very significant, the correlation is not as high as the experiments for the MPs dataset in Section 3.4.2, which can be attributed to a variety of factors. Initially, given the semantics of the two methods, it is not meaningful to directly compare them because the arrangement with the MP projections displays clustering correlation while the arrangement of the news sources shows a measure of precise arrangement in a linear axis, which is by its nature a more difficult problem. Moreover, it is possible that the minimum cost arrangement of the news sources might not correspond exactly to the left-to-right political spectrum because the dynamics of the Twitter network are very complex and heterogeneous among its users.

The DeGroot Model Approach

The application of the DeGroot Model Approach on the enriched projections aims at the extraction of the political affinity and orientation of the news sources. The NOIs of the enriched projections are handled using the same guidelines of Section 3.4.2. Each undirected edge is duplicated into two opposite directed arcs and the weights of the outgoing arcs of every node are normalized. Furthermore, seven additional nodes are introduced that represent the political parties and the MPs are linked to their corresponding party. We note that the nodes of the news sources are not directly connected to any party but their connections with the MPs are the indirect link with the parties that we aim to examine and evaluate.

We transform the nodes of the MP's into stubborn agents (i.e., nodes that are solely influenced by their party and their connections to other MPs are ignored) and we estimate the political affinity of each news media node by evaluating the hitting probability of a random walk that originates from it to each party's node. The retrieved probabilities are then divided by the corresponding parliamentary group's size, in order to avoid any dominance effect by the largest groups.

A series of preliminary experiments prove that the addition of the 24 nodes of the news sources to the graph does not taint the validity of our approach. We apply the DeGroot model in order to classify each MP to a party (using the same methodology as

Table 3.10: Correlation coefficients (Kendall tau-b and Pearson), P@5 and P@10 values of the news media rankings for all parliamentary groups of the DeGroot experiment compared with the experts' survey.

Ranking Probability	τ_B	ρ	*P@5*	*P@10*
SYRIZA	0.57063	0.76702	0.8	0.8
ND	0.51756	0.81495	0.2	0.6
DHSY	0.63157	0.75014	0.8	0.6
POTAMI	0.58757	0.75780	0.8	0.6
ANEL	0.11213	0.03983	0.2	0.5
XA	0.57366	0.64034	0.8	0.7
EK	0.31555	0.57076	0.4	0.5

in Section 3.4.2) and collate the results in Table 3.9. The hit ratios on the classification of the MPs to their parties are slightly decreased in most cases while in the case of the graph produced by the overlap projection the ratio is increased (from 87.65% to 88.27%). These encouraging results allow as to proceed to the evaluation of the political affinity and orientation of the news sources.

Political affinity extraction The political affinity extraction of the news media to each party is achieved through the calculations of the hitting probabilities for random walks that originate from the news media nodes to the party's node. Our hypothesis for this experiment states that the news media that share common ideological and political views with a specific party should also share common followers with it and, thus, their links to the party's MPs in the projection graphs should be strong and would result to an increased hitting probability of a random walk that originates from the news media node to the specific party's node.

We test our hypothesis on the social graph produced by the overlap projection since our preliminary experiments suggest that the addition of the 24 news media NOIs enhance the mechanism of the DeGroot model in the extraction of political information. The experiment produces a 24×7 matrix that contains the uniform per party influences of each party to the news media node. We round these influences to 2 decimal places to produce coarse grain results and avoid any jitter.

In order to evaluate the validity of our approach, we correlate the results of our experiment with the results of the political affiliation from the expert survey. More specifically, we produce separate rankings of the news media according to their influences to the nodes of all the parliamentary groups in our dataset and correlate them with the findings of the experts' survey. We assess the correlation using the Kendall tau-b (τ_B) and the Pearson correlation coefficient (ρ). The results presented in Table 3.10 validate our approach. The values of the Kendall tau-b coefficients are in most cases greater than 0.5. The results for ANEL and EK (the two smallest parliamentary groups in the graph) are significantly lower, a phenomenon that could be attributed to the small number of corresponding NOIs that are included in our dataset (4 and 2 NOIs respectively). In general, the findings provide a clear indication of dependence between the produced

rankings and the experts' survey results. Furthermore, the high values of the Pearson coefficient ρ in almost all cases confirm our hypothesis that our method reveals information about the political affinity of news media from our dataset.

In order to make these results more easily understandable to a broader audience, we utilize the metric of *Precision at K (P@K)* to calculate the precision of the top K elements in the produced rankings in respect with the top K elements of the findings from the experts' survey. The results presented in Table 3.10 further confirm the correlation of the produced rankings with the experts' survey results. The values of *P@5* are surprisingly high in the cases of SYRIZA, DHSY, POTAMI and XA while the low values of ND, ANEL and EK can be attributed to the small number of NOIs in the dataset (in the cases of ANEL and EK) and to divergence of opinions between the political scientists and the Twitter users about the news media that are the greatest supporters of ND. The values of *P@10* exhibit significant consistency ranging between 0.8 and 0.5 providing, thus, further support to our findings.

Political orientation extraction We further extend our approach and alter our experimental scenarios to extract information about the political orientation of the news media from our dataset using the DeGroot model. We modify our produced social graph with stubborn agents by discarding the nodes of 5 parties (the nodes of ND, DHSY, ANEL, POTAMI and EK). The remaining 2 party nodes (SYRIZA and XA) are considered as the two poles of the left-right ideological spectrum, according to their self-identification, that are present in our dataset. We evaluate the hitting probabilities of random walks that originate from the news media nodes to the 2 parties' nodes. Our hypothesis is that the MPs' location in the political spectrum is reflected in the arcs of our social graph and, thus, the news media links to the MPs result to high hitting probabilities to the pole of the political spectrum that are closer to.

The results of the experiment produce a ranking of the media based on their "distance" from the pole of SYRIZA (i.e., the party that is considered to represent the leftmost pole in our political spectrum). The ranking is then correlated to the findings of the second question from the experts' survey concerning the political orientation of the news media. The Kendall tau-b coefficient is evaluated to 0.58792, a result that indicates strong correlation between the two rankings and validates our hypothesis, while the Pearson coefficient is evaluated to 0.46833 that further supports our approach.

Figure 3.8 presents graphically the rearrangement of the news sources in the results of the DeGroot approach compared to the ground truth of the experts' survey. A prominent observation is that the rankings are in agreement about the locations of the sources that are closer to the poles (i.e., the leftmost and rightmost edges of the stripes). Furthermore, the differences in the arrangement of the news sources in the center of the political spectrum (i.e., the middle of the stripes) are noticeable but not extensive.

(a) Stripe produced by experts' survey results.

(b) Stripe produced by MinLA algorithm (overlap).

(c) Stripe produced by DeGroot experiment's results (overlap).

Figure 3.8: Comparison of the three rankings produced by the experts' survey, the MinLA experiment and the DeGroot approach. Stripes (a), (b) and (c) consist of rectangles that visualize the ranking of the news media. Each rectangle represents a news source and is coloured according to the news source's affinity to the two poles of the political spectrum based on the results of the experts' survey (violet and yellow for the left and right pole respectively). The resulting colours are preserved in stripes (b) and (c) for comparison purposes. Due to its nature this figure might not be readable in grayscale form.

3.5.4 Discussion

The purpose of the news media case study was to examine if our methodology can be applied to the enriched dataset to determine the political affinity of news sources. Overall, we have showed that modularity clustering, the MinLA problem and the DeGroot model are suitable methods. The results were evaluated against the replies of the expert survey and indicate that these methods can be used to study the news sources and determine their political affinity with significant precision.

While the accuracy of the methods in both Section 3.4 and Section 3.5 were very high, the evaluation over the MPs was finer and almost flawless. A possible explanation is a fundamental distinction among the MPs and the news sources as Twitter users in the context of our study. In particular, it is an established fact that the dominant act of politicians in Twitter is political deliberation and advocacy and, thus, it is reasonable to assume that other users follow them because of their political standing. In Section 3.4, we have proved this assertion with significant accuracy by deriving the political standing of MPs via their followers. However, this does not always seem to be the case for the news sources. News sources convey perspectives of online presence other than politics and, as a result, users may follow them for reasons unrelated to politics, for example sports news. This fact is a form of interference on our methods which rely on the assertion that users follow NOIs for reasons related to the context (politics in this work). Filtering out these users is outside the scope of this study but we believe it would improve the accuracy of the method even more.

Finally, in Figure 3.8 we provide a visual juxtaposition between the news media rankings of the MinLA algorithm and the DeGroot model for the overlap projection. The two

methods achieved similar τ_B evaluation (0.6249 and 0.5879 respectively) against the experts' survey ranking. An apparent similarity of the arrangements is the placement of the left wing (violet). While the two arrangements display similarities, it can be concluded that the MinLA ranking has more matches in the center of the spectrum while the DeGroot ranking has more matches in the right end.

3.6 Conclusions

The purpose of this work was to study and assess the possibility of deriving political affinity of particular entities (NOIs - Nodes of Interest) using the Twitter follower network. We initially applied our methodology on the Members of the Greek Parliament in order to a) classify them in political parties, b) arrange them in a bipolar spectrum and c) determine their correlation factors with political parties. Our results suggest additional evidence about the validity of the hypothesis that Twitter followers can portray the political leanings of their followees. Our work was later extended on the enriched dataset containing the MPs as well as popular news sources that operate under a political context.

Overall, our approaches are simple to implement and easy to reproduce while delivering very significant accuracy. Furthermore, the overlap coefficient, an underutilized measure, especially in the context of social networks, is highlighted and shown to achieve a considerably superior efficiency compared to other projection functions. Additionally, we applied concepts to this problem that have not been examined in prior literature, the MinLA problem and the DeGroot model with the presence of stubborn nodes, and showed that these techniques are perfectly suitable for the analysis of our dataset. We deem that the application of these novel ideas will have a theoretical impact on future research regarding Social Network Analysis. Our methods work without any prior knowledge of the ground truth related to the NOIs and do not employ heavy preprocessing or filtering on the raw data.

We argue that the proposed methodology could be utilized in other practical situations. While the examined scenarios of this work focus on the political attitude of the NOIs, a possible application of the methods could be targeted for other interest domains. For example, the tourism industry is a domain with extensive presence in online social networks. The analysis of online social networks' structure and nodes under the prism of their touristic interest could unveil beneficial aspects of their behaviour and provide valuable findings to tourism organizations.

There are several lines of research arising from this work which should be pursued. A natural extension of this work is the investigation of the scalability of these methods when applied to other countries and, consequently, different political systems. The political scenes among different countries are very diverse and it is interesting to see if the methods are suitable and under which settings. Moreover, the dataset used in this

work can also be enriched with temporal information regarding the establishment and possible cancellation of followerships. Therefore, various observations can be resulted from a temporal study during critical points in time, such as elections. Finally, an interesting emerging topic in the field of graph applications is graph embedding (Goyal & Ferrara, 2018), where vertices of the graph are represented in vector space. The application of graph embedding in our dataset, and possibly an extensive comparison with the results of this chapter, could be addressed in future studies.

Chapter 4

Application: Point of Interest Lists in Recommendation Systems

Location based social networks, such as Foursquare and Yelp, have inspired the development of novel recommendation systems due to the massive volume and multiple types of data that their users generate on a daily basis. More recently, research studies have been focusing on utilizing structural data from these networks that relate the various entities, typically users and locations. In this work, we investigate the information contained in unique structural data of social networks, namely the *lists* or *collections* of items, and assess their potential in recommendation systems. Our hypothesis is that the information encoded in the lists can be utilized to estimate the similarities amongst POIs and, hence, these similarities can drive a personalized recommendation system or enhance the performance of an existing one. This is based on the fact that POI lists are user generated content and can be considered as collections of related POIs. Our method attempts to extract these relations and express the notion of similarity using graph theoretic, set theoretic and statistical measures. Our approach is applied on a Foursquare dataset of two popular destinations in northern Greece and is evaluated both via an offline experiment and against the opinions of local populace that we obtain via a user study. The results confirm the existence of rich similarity information within the lists and the effectiveness of our approach as a recommendation system. The results presented in this chapter are published in Stamatelatos et al. (2021).

4.1 Introduction

User generated content and recommendation systems. Point of Interest (POI) ratings and recommendations are valuable to tourists and enable them to explore new places to visit. It is, therefore, no surprise that there is plenty of active research on recommender systems that, using data from various sources, are able to make personalized venue recommendations. Traditionally, these systems have been relying

on text, image or other multimedia content, but with the rapid development of online social networks (OSNs) and location based social networks (LBSNs) (Zheng & Zhou, 2011a), such as Foursquare and Yelp, link analysis based approaches have gained significant popularity. These LBSNs acted as catalysts to provide an abundance of various forms of data to either inspire the development of novel recommendation systems or enhance existing ones (L. S. Jackson & Forster, 2010). Typically, this data originates from the actions of the network users themselves and, thus, are often called user generated content (UGC) (Lu & Stepchenkova, 2015).

Link analysis based methods for recommendation usually come in the form of explicit or implicit graph structures. In a recommender system where users are being recommended items, UGC consist of the user-user or user-item structures, or a combination of these. For example, the user-user layout might refer to the similarities in the behavior of users, while user-item might be explicit relations among users and items, such as check-in history.

Although there has been extensive research that is based on user centric (user-user and user-item) link schemes, such as item ratings, GPS trajectories as user-location network, check-in data and similarity among users, there has been a lack of research on link analysis based recommender systems that use an underlying item-item form of relational structure. In this work, we utilize a common feature of social networks and LBSNs, namely the *lists*, to create a collection-item bipartite graph structure and utilize that to infer an item-item topology of POI similarities. These similarities are then used to produce personalized POI recommendations for users by also taking into consideration their profiles.

Foursquare lists. We make use of the POI lists information from the Foursquare LBSN, an online portal with rich user generated information, but our method does not rely exclusively on this provider. Our approach is being developed on the plausible assumption that lists are collections of related POIs, an assertion that can be attributed to the features and properties of Foursquare lists. The Foursquare lists feature was implemented in 2011 in order to allow users to keep track of the places they have been or discover places they are willing to go. The lists can be made quickly from a user's check-in history or they can be created from scratch while users can share lists with friends, follow and contribute to public lists. Essentially, this feature leverages the users' intent to visit places to enable *future check-ins* (or To-Visit lists), a generalization of physical check-ins that is not bound by the users' location history.

Users, naturally, utilize lists for multiple purposes too, but the pattern is that POIs that are listed together are more likely to carry similarities. Another use of lists is for opinionated best places to visit recommendations, often for specific regions, for example "Best places to visit in Chalkidiki". Lists can also portray the favorite places of a user or advertise a certain category of POIs, for example "Sea sports in Kefalonia". The lists

contain, in their largest portion, places of entertainment, such as cafes and restaurants, but they also include POIs which fall into different categories, such as landmarks and archaeological sites. Moreover, Foursquare lists capture one more dimension than the traditional check-in history information since users tend to create lists of their check-in histories based on different criteria, usually temporal, spatial or both, for example "Santorini, Summer 2019". This is a further indication that POIs inside a list are related to each other as they are organized based on some criteria that users find reasonable.

Overall, we argue that the users who created the lists, or contributed to them, intentionally grouped together a list of POIs which, in turn, implies that these POIs are, by at least one relevance measure, related to one another. As a result, the coexistence of two POIs in lists is a measure of the likelihood that one user who likes one of the them will like the other one too. The users may be considered as entities that produce sets of relevant POIs according to their own unknown judgement, while our method aims to leverage this phenomenon in order to extract useful information for POI recommendations.

Our proposed method. Point of Interest lists, and the assumption that POIs in the same list are related, is the building block of the method that we propose in this chapter. Our approach uses this information along with the profile of a user, that specify their preferences with respect to the recommendation context, to offer personalized POI recommendations that correspond to the best places to visit for that particular user. The list data structure can be seen as a bipartite list-location graph, where edges are interpreted as the containment of a POI in a list. Using this bipartite structure, we generate a pairwise similarity matrix of all the POIs, a process that is called *one-mode projection*. Determining, however, when a POI is similar to another may be subject of personal preference and there exist multiple perspectives of how similarity is imprinted in the POI lists. Thus, we use various different weighting methods to generate the similarity matrix and assess the effectiveness of each one. Weighting methods often appear in literature about *link prediction* (Liben-Nowell & Kleinberg, 2007; Zhou et al., 2009) or *proximity* (Goyal & Ferrara, 2018). Our algorithm, then, analyzes the profile of the user and assigns a relative score to every POI based on their similarity with the user preferences. The final personalized recommendations that are given are considered the POIs with the greatest relative scores. We finally evaluate our approach against a Foursquare dataset in two popular tourist destinations in northern Greece via both an offline experiment and by performing a user survey to obtain the ground truth for the experiment from inhabitants of these areas. At the same time, we also assess the various weighting methods to uncover their properties and highlight their differences.

Regarding the properties of the proposed approach, we argue that utilizing the Foursquare lists for recommendation systems has some inherent advantages in regards to the methodology and the quality of the results. Initially, as a type of information, Foursquare lists create a richer source for personalized POI recommendation since

people often want to visit more places than they actually do. In typical user-location methodology schemes, where this information comes from the check-in history, it is not taken into consideration where the user will check-in in the future. Furthermore, often, these types of user-location graphs capture only part of the information, whereas on a list, POIs are arranged and grouped using specific criteria, usually temporal. Another limitation of check-in history systems is that a user's visit to a location does not necessarily imply that the user liked it. As a result, the check-in history (a single list) might contain POIs that are both interesting and uninteresting to the user and, hence, its contents cannot always be considered related. On the contrary, the action of submitting a POI to a list is more conscious and can even be corrected later since the user can remove the POI from that list.

Moreover, an inherent advantage of the proposed method is the ease of applicability to the new users of the network. The data via which the similarities among the POIs are inferred are already publicly available and, thus, a new user only needs to have a profile stating their preferences to receive personalized recommendations. In particular, a new user can easily be introduced into the proposed system, without the need for their friends to participate or a huge user base to be available, as is often the case in collaborative filtering approaches. This property is referred to as *cold start* (Sertkan et al., 2019), a situation related to a common problem in recommendation systems: the system may not be able to generate recommendations until a significant amount of information has been gathered. Despite not being sensitive to new users, our approach is sensitive to new items as new POIs that appear in the LBSN are required to establish list connections before being incorporated into the pool of potential recommendations. This process is expected to take some time, depending on the popularity of the POI and the amount of users engaging in that context. However, the list structure construction does not require any additional effort as the users are already creating and maintaining lists for their own benefits while, at the same time, they are submitting useful information that can be leveraged for suggestions to other users as well.

Lastly, the use of Foursquare lists does not raise any privacy or royalty issues since it is publicly available and accessible via the Foursquare API. For example, the check-in history that is the core of multiple approaches to recommendation is considered personal data while a list might express the same or more information without exposing sensitive data. The approach we describe in this chapter is based on purely structural data and it is worth noting that we treat Foursquare lists simply as their ID and no additional information is required, such as the user who created that list or its name. This further extends the field of applications to domains where this information is very difficult to acquire or not available at all, for example where lists are implicit. In addition, our method is easy to reproduce and can be performed at real time because the similarity matrix among the POIs can be computed ahead of time and the main component of the recommendation is a very simple mathematical operation, a weighted sum.

Contribution. Our contributions can be summarized as:

- We highlight the aspect of lists of POIs: a generalization of check-in history and a richer source of information for personalized POI recommendation systems.
- We argue that lists of POIs, as primarily information provided by users, are considered *User Generated Content* and should be more openly addressed in the literature of tourism applications.
- A rich Foursquare dataset of POIs and POI lists that was acquired over a period of several months is shared with this work.
- The attainment of considerable improvements over the existing modified MI methodology via a comprehensive study of multiple similarity measures, which are also contrasted in the context of tourism.

Outline. The rest of the chapter is organized as follows. In Section 4.2, we explore recent literature about recommendation systems based on LBSNs, graph processing recommendation and entity similarity approaches. In Section 4.3, we present the Foursquare dataset of POIs and POI lists that was created as part of this work. Our methodology and the representation of the POI list dataset are presented in Section 4.4. In Section 4.5, we evaluate the system and demonstrate the existence of rich information within the Foursquare lists and the effectiveness of our method as a recommendation system. In Section 4.6, we discuss the concept of recommendation diversity and how our method appears under this prespective. In Section 4.7, we discuss several subjects related to the information that is embedded in the POI lists and the relationships among the similarity measures. Finally, Section 4.8 concludes this chapter and presents suggestions for future work.

4.2 Related Work

A first attempt that utilizes lists of POIs for recommender systems was presented in Karagiannis et al. (2015). In that work, the similarities among POIs are approximated based on a modification of Mutual Information (MI), while in this chapter, we make a complete recommendation system and evaluate it against local inhabitants' opinions. The current system is, furthermore, capable of incorporating user preferences for personalized suggestions and, at the same time, we enrich the similarities with more weighting functions, both set theoretic and graph theoretic, and show that they can outperform the modified MI measure. Moreover, the results among the similarity measures are juxtaposed and useful conclusions are drawn from these that determine the properties of each similarity function. In the rest of this section, we present previous work that is related to the views expressed in this chapter.

The role of LBSNs in recommendation. Developments on mobile devices and the emergence of more advanced online tourism portals has provided researchers with motivation and valuable data to support the investigation of POI recommendation based on LSBNs. Current state of the art is given in Bao et al. (2015), Ravi and Vairavasundaram (2016) and Eirinaki et al. (2018) and the references therein. In general, POI recommendation systems can be categorized by methodology as content based, link analysis based and collaborative filtering based.

The method described in this chapter constitutes a link analysis method of standalone personalized POI recommendations, where the links refer to the explicit relations among POIs and lists. In particular, we infer a POI-POI network based on this bipartite POI-list network. This structure appears to be a less studied field in the literature, where typically the links refer to user-user or user-POI relations. Several of the recommender systems in the literature can also be applied to a context other than tourism, such as movies, shopping items and books. Similarly, they can be utilized for user recommendations, for example to suggest "people you may know", "people to follow" or activity recommendation, such as sightseeing, boating and jogging.

Recommendation using graph processing. The existence of explicit or implicit links in social networks has motivated the use of graph processing and graph theoretic approaches for POI recommendation.

In Wu et al. (2015), the authors studied options of clustering POIs and users based on information from geographic social networks. In particular, they considered the *social distance* between two POIs as the Jaccard index of the users that have checked-in in those POIs and leverage this index to partition the POIs into groups of similar places. In a recommendation perspective, the fact that two commercial places belong to the same cluster indicates that there is a high likelihood that a user who likes one place will also be interested to visit the other.

Two datasets containing geolocation and temporal information concerning users in 11 cities were utilized in Noulas et al. (2012) in order to evaluate a proposed recommendation algorithm in comparison to several known techniques. The methodology was based on a random walk in the graph of users and POIs, on which the edges represent friendships among users or check-in actions among users and POIs, while the recommendations were ranked according to the hitting probabilities of the random walk. The authors mentioned that all versions of collaborative filtering in their experiments, that were supposed to better model users' preferences, fail to outperform the popularity based baseline.

In Wang et al. (2013), the authors suggested a recommendation algorithm that operates using different factors: a) past user behavior (visited places), b) the location of POIs, c) the social relationships among the users, and d) the similarity between users. By

analyzing the publicly available data of Gowalla[1], they showed that more than 80% of the new places visited by a user are in the 10km vicinity of previous check-ins and more than 30% of the new places visited by a user have been visited by a friend or a friend-of-a-friend in the past. These facts imply that geographical and social information significantly affect the choices of a user when deciding which new place to visit. Therefore, it is also desirable for a recommender system to take these components into consideration.

Moreover, in Kefalas et al. (2018), a recommendation system was proposed that incorporates user time-varying preferences as well. In particular, the recommendation was based on a tripartite graph, consisting of users, locations and *sessions*. Sessions can be considered as a more specialized form of POI lists, where the criterion of POI coexistence in the same session is temporal. The authors concluded that the time dimension plays a very important role in recommender systems.

Finally, an interesting work that aimed to identify and validate the heuristic factors affecting the popularity of "best places to visit" recommendations was presented in L. Li et al. (2019). This empirical study focused on the explicit best places to visit listings in Qyer.com, a concept that is also a subset of Foursquare lists which contain opinionated recommendations ("Best places to visit in ...").

Similarity concepts in tourism. Often, similarity measures are utilized in order to express the relations among entities in a link-based system, when these are not explicitly present. For example, in Celik and Dokuz (2018), the researchers utilized data from Twitter to extract the check-ins of users and proposed a methodology for revealing the socially similar users based on their online traces. A bipartite network between tourists and POI reviewers was used in Ahmedi et al. (2017) to drive a collaborative filtering approach to recommendation. The user-user similarity weights were inferred using the Jaccard index and the cosine similarity, while the method was evaluated on a Foursquare dataset. Another establishment and tourism portal, TripAdvisor, was utilized in Van der Zee and Bertocchi (2018) as a medium of social network analysis application on the POI reviews. An indirect similarity matrix of POIs was created based on the two-mode network of users and reviewed POIs; the similarity between two POIs was defined as the intersection cardinality between the users that reviewed these POIs. In this chapter, we heavily utilize the notion of similarity and we argue that there are multiple perspectives as to what constitutes similarity. As a result, we propose multiple measures to express this similarity among POIs and perform extensive experiments to compare the properties of those functions.

The notion of similarity among entities can be extended, besides POIs or users, to other layers as well. For example, in Preoţiuc-Pietro et al. (2013), the concept of city similarity was studied, where each city is represented by the collection of POIs in it. In particular,

[1]A LBSN that operated until 2012, primarily under its mobile application.

a city is represented as a vector of the categories of the POIs within. Another form of similarity was discussed in Sertkan et al. (2019), the attribute similarity, which defines the pairwise relation among different tourism attributes, such as island, mountains, river, family, diving and others. Finally, in David-Negre et al. (2018), similarities among tourism domains (TripAdvisor, Booking, Trivago) were approximated; these similarities denote whether they have been used by the same tourists.

POI sequence recommendations. An interesting concept in the field of POI recommendations is the conception of trip planning, which refers to the recommendation of a collection of POIs that are bound by a common characteristic. Such systems can be useful for trip planning as the recommendation describes a collection of venues for a trip or a route. Some POI group recommendation systems were the subject of recent papers (Cenamor et al., 2017; Rakesh et al., 2017; Wörndl et al., 2017; Arentze et al., 2018). Trip planning has an indirect connection to POI lists because a trip, as an abstract set of POIs, can be thought as a POI list with the characteristics posed in this chapter: a collection of related POIs.

Another interesting concept is the session-based recommendation approaches which are recommendation techniques that aim to predict the user's immediate next actions (Quadrana et al., 2018), such as next-item recommendation (Song et al., 2015; Hidasi et al., 2016) and list continuation (Hariri et al., 2012). In Ludewig and Jannach (2018), the authors present the results of an in-depth analysis of a number of complex algorithms, such as recurrent neural networks and factorized Markov model approaches, as well as simpler methods based on nearest neighbor schemes. Their results indicate that the simpler methods perform equally well as more complex approaches based on deep neural networks.

In such systems, it is a common technique to use offline evaluation schemes. In the domain of next-track music recommendation, often only the last element of a sequence is hidden, while in the recommendation of videos in streaming platforms (Hidasi et al., 2016) an approach is taken where the number of hidden elements is incrementally increased. In the case of next-track music recommendation (Hariri et al., 2012), the authors use leave-one-out cross validation since they do not take into account the creation time of the playlists. In this work, we also utilize leave-one-out cross validation for the evaluation of our method during the offline study.

Recommendation using graph embeddings. In Mighan et al. (2019), the authors propose a graph-based POI recommendation approach that utilizes users' check-ins at specific time points in social networks, modelling them in a heterogeneous graph. Their approach is supported by a neural network embedding method and the efficiency of this method is evaluated using Foursquare dataset providing improvements over other existing graph embedding approaches. Graph embedding approaches are also proposed in other works, such as Christoforidis et al. (2018), where the embedding is

Prefecture	POIs	Occurrences in lists
Chalkidiki	7,590	14,291
Cephalonia	2,612	3,671
Cyclades	2,171	19,489
Dodecanese	5,524	10,160
Drama	1,077	245
Kavala	4,002	11,129
Serres	1,970	400
Thessaloniki	21,266	55,532
Xanthi	1,533	391
Total	47,745	115,308

Table 4.1: Dataset summary as the number of POIs in each prefecture and their occurrences in lists.

applied on unipartite and bipartite graphs.

4.3 Dataset

For the purposes of our project we retrieved and assembled a dataset of POIs and lists from Foursquare covering a large geographic portion of Greece. In particular, the dataset contains 47,745 POIs located in 9 prefectures of tourist interest and 17,000 lists with 115,308 mentions of the POIs. The dataset exhibits a wide diversity of categories regarding the POIs (e.g., restaurants, hotels, car-rentals, archaeological sites, shopping malls, gyms, rivers, churches) while the covered area includes popular touristic destinations (e.g., Rhodes, Santorini, Chalkidiki). The geographical breakdown of the POIs is presented in Table 4.1. The complete dataset contains data that were retrieved until October, 2020, and is available online[2]. The dataset can be continuously augmented as long as the retrieval process is performed since new POIs are added in Foursquare and its users create new lists every day.

A temporal examination of the lists in Foursquare reveals interesting characteristics about the users' behavior. Figure 4.1 presents the histograms of the per month list creation and update dates from 2011 to 2020. It is more than evident that the two distributions bare great resemblance since the vast majority of the local maxima follow the same pattern. The maxima correspond to each August of the respective year, a month that traditionally demonstrates a peak in tourism attention. The years 2017, 2018 and 2019 contain the majority of lists with 52% created and 57% updated during this period while 2020 exhibits a significant drop in generated content, possibly related to the COVID-19 pandemic.

The information contained in the two date fields (i.e., the date of creation and the date of the latest update) might potentially be very useful, e.g., for filtering very old lists

[2]https://doi.org/10.17632/dx39jx9v5p

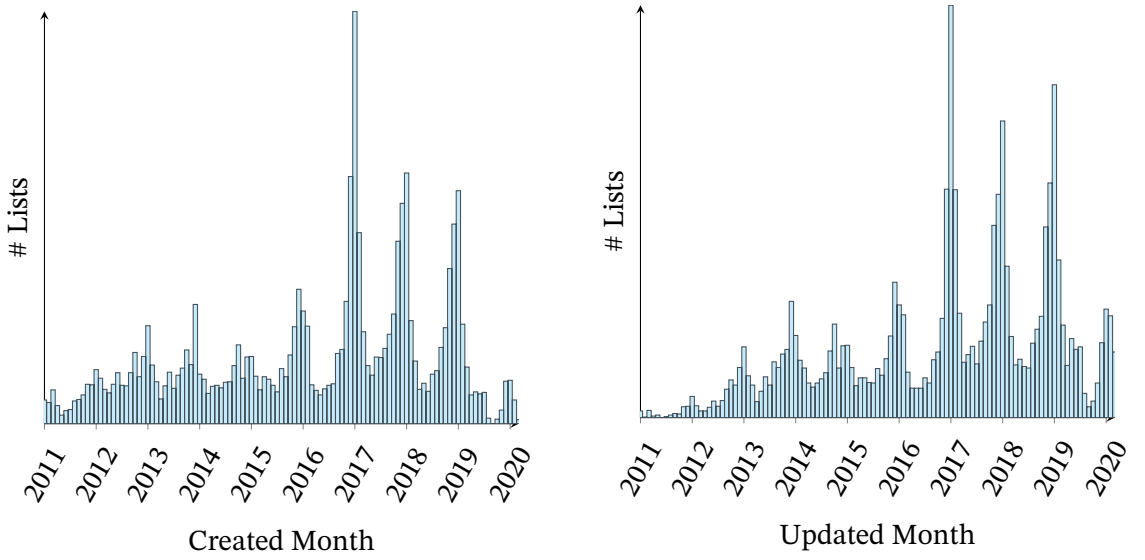


Figure 4.1: Created (left) and updated (right) dates distribution of the POI lists. Each bar corresponds to one month from August 2011 to October 2020.

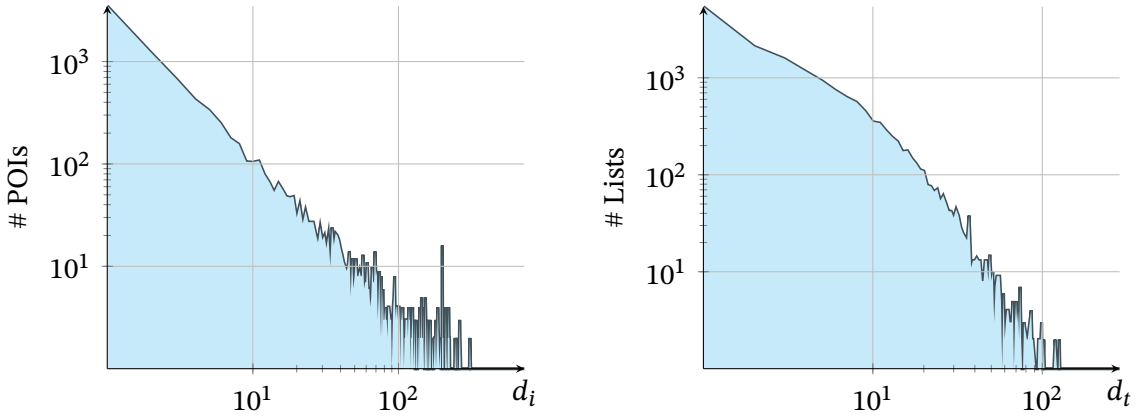


Figure 4.2: Log-log density distribution of occurrences for POIs (left) and lists (right) in the full dataset.

which could be outdated. In our work, we use the dataset as acquired by the Foursquare API in order to simplify the process and demonstrate the potential of lists without preprocessing or filtering. We argue that most of the dates in the dataset are up-to-date since the majority of the lists were either created or updated recently. Despite this, utilizing this field, for example to filter the most recent lists, could be pursued in the future, possibly in combination with other list metadata, for example their title or their authoring user.

Further analysis of the dataset highlights interesting properties that support our incentive to utilize Foursquare lists in recommendation systems. Figure 4.2 presents the log-log plot of the occurrences distributions for the POIs and lists in the dataset. The plots provide solid indications about a power law behavior of the distributions, a feature

Area	POIs	Lists	Occurences in lists
Thessaloniki	2,871	6,534	49,448
Kassandra	526	1,286	6,336
Total	3,397	7,820	55,784

Table 4.2: Summary of the case study dataset as the number of POIs and occurrences in lists.

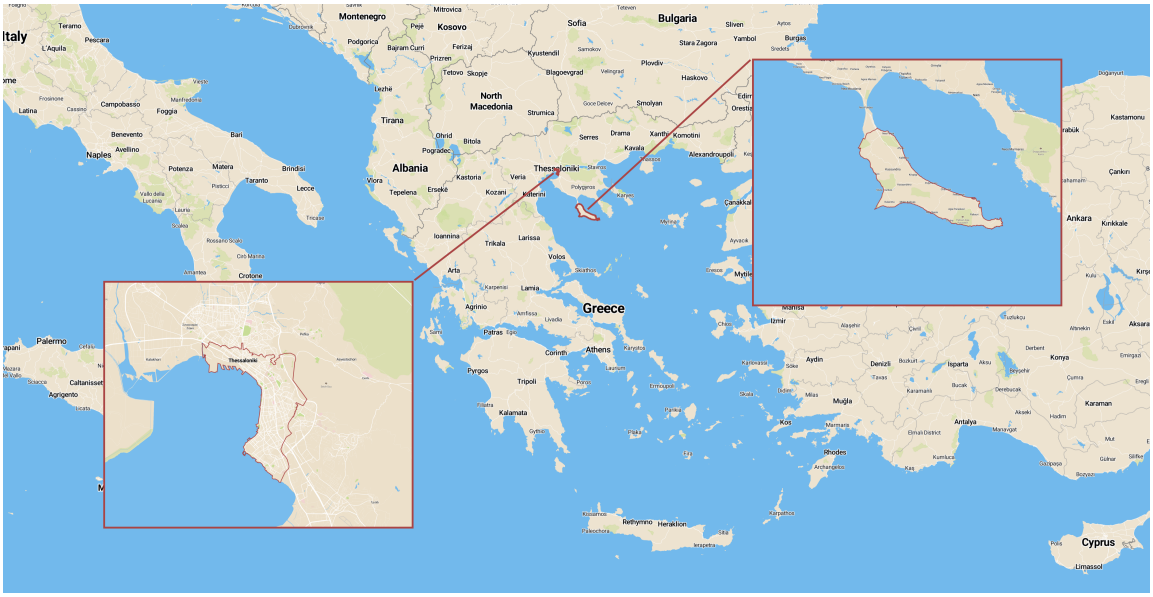


Figure 4.3: Map of Greece showing the two areas of our dataset where we perform the case study experiment.

that is commonly observed in datasets of recommendation systems (Goel et al., 2010; Abdollahpouri et al., 2017; Belletti et al., 2019). This phenomenon can be attributed to the fact that people tend to perform additive actions on POIs (i.e., insert into list) with a higher rate for POIs that are already popular (Clauset et al., 2009).

In this study, we examined the performance of our methodology in a case study using a preliminary subset of the complete dataset that was assembled until February, 2020, and focused on two cities with vivid touristic activity, namely *Thessaloniki* and *Kassandra*. These two areas are in northern Greece, contain multiple types of leisure and entertainment attractions and are both established as notable tourist destinations; Thessaloniki is the second largest city in the country, and Kassandra is the most visited area of the suburban region of Chalkidiki. Our case study dataset contains 2,871 POIs in Thessaloniki and 526 in Kassandra and 7,820 lists with 55,784 mentions to POIs forming, thus, a subset of the complete dataset. The geographical breakdown of the data is summarily presented in Table 4.2 while Figure 4.3 shows the geographic areas that are covered by the case study.

Analytical findings on the preliminary assembled dataset confirm the consistency of certain characteristics with the augmented complete dataset. Figure 4.4 presents the

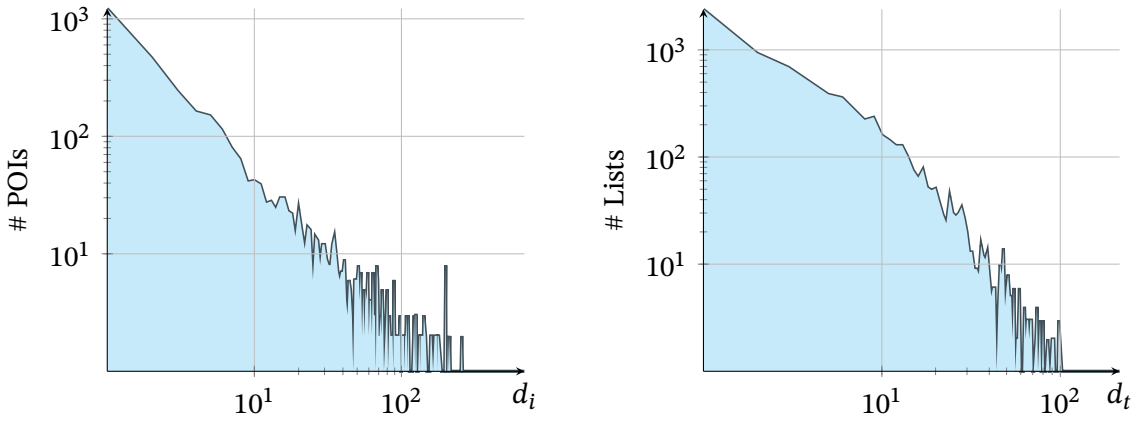


Figure 4.4: Log-log density distribution of occurrences for POIs (left) and lists (right) in the case study dataset.

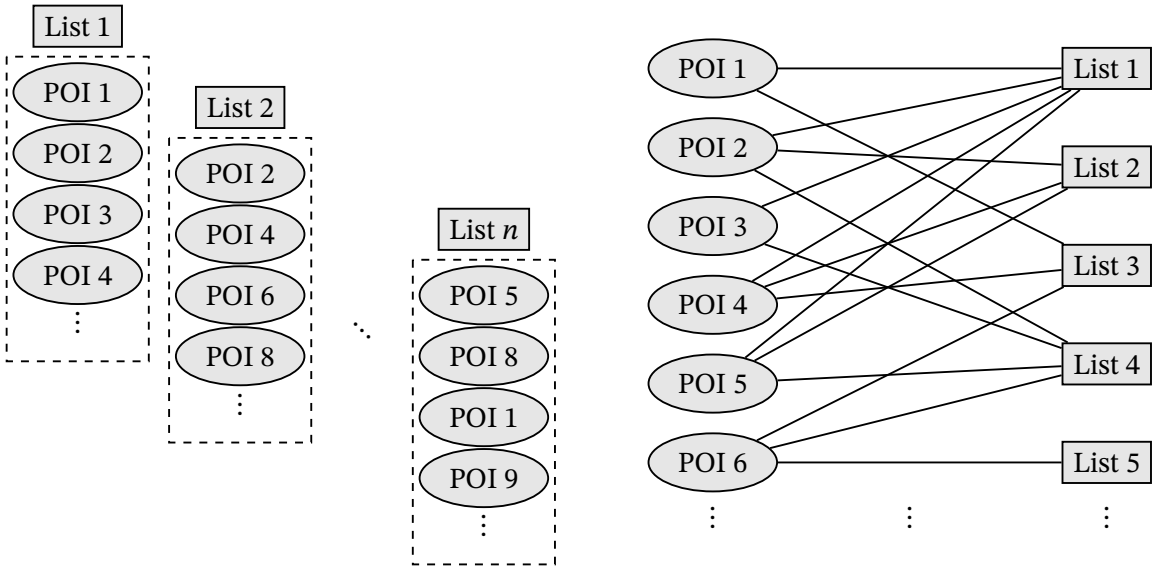
log-log plot of the occurrences distributions for the POIs and lists in the data subset. A prominent observation of the plots, when compared with the respective plots of the complete dataset in Figure 4.2, is that the data exhibit a remarkably similar behavior that resembles a power law distribution and indicate a scale-free property. This is something to be expected because people tend to perform additive actions on POIs (like, insert into list etc) with a higher rate for POIs that are already popular. It is also generally observed that distributions across users and items (in this case POIs) exhibit power law behavior (Belletti et al., 2019, Section IV).

In fact, an approximation of the maximum likelihood estimation method given in M. Newman (2018, Equation 8.6) yields an exponent of 1.72 for the POIs and 1.79 for the lists[3]. Hence, we can safely argue that our case study is performed based on a representative subset of the data and the evaluation of our methodology on this subset provides credible insight of its performance on the complete dataset. This property is a preliminary positive indication about the sufficiency of the quantity and quality of the available data. However, a more thorough investigation of this issue would be an important future work direction and would facilitate further applications of our approach.

4.4 Methodology

In this section, the method of the personalized POI recommendation system is presented. As stated previously, the system uses the Foursquare lists as the only source of information. It requires a user profile as a preference vector of POIs and is capable of quantifying the preference score of an arbitrary POI that is not present in the profile. Thus, the method is able to predict a relative score that the user would assign to this POI,

[3]More exactly, 1.7249867410600555 for the POIs and 1.7883234540454382 for the lists and with expected statistical error σ 0.012438908998926225 and 0.00904507078192265 respectively.



(a) POI list collections. **(b)** Bipartite graph representation.

Figure 4.5: (a) Abstract example of lists as collections of related POIs and (b) its representation as a Lists - POIs bipartite graph, where an edge represents the existence of a POI in a list.

or otherwise how likely they are to like the POI with respect to other POIs. This method can operate as a personalized recommendation system by assigning relative preferences to all candidate POIs in the context and selecting the ones with the maximum values. On another perspective, it can also be used as a negative recommender to suggest POIs least related with the user profile, for example POIs where the user is not recommended to visit.

This section is organized based on the components that the method consists of. In Section 4.4.1, the structure of the list dataset and its representation as a bipartite graph is presented. The similarities among the POIs are then approximated in Section 4.4.2 using various similarity functions. In Section 4.4.3, the user profile is introduced into the system while the main component of the method that assigns relative scores to POIs is explained in Section 4.4.4.

4.4.1 Dataset Representation

The dataset that drives the approach consists of the Foursquare user generated lists that refer to the geographic areas of interest, which are the areas that the recommender system is to be deployed and operated. The extent of the lists also corresponds to the geographic extent of the operation of the recommender, such as a city, a administrative prefecture, or even a whole country. Every list in the dataset can be portrayed as a collection of one or more POIs, such that the same POI may exist in multiple lists while there could also exist multiple identical lists (with exactly the same POIs).

A natural representation of the lists is a bipartite or two-mode graph, where one set

of vertices is the lists and the other set of vertices is the POIs. List x and POI y are connected via an edge if and only if y belongs in list x. An abstract example of the POI lists structure as well as its representation as a bipartite graph is given in Figure 4.5. Each list is only described by an arbitrary ID and no additional meta-information is required, while each POI is a named location and refers to a specific business or attraction.

4.4.2 POI Similarity Matrix

The similarities among the POIs in the dataset are, then, estimated using the lists, based on the assumption that every list is a collection of related POIs. This relation among the POIs defines the likelihood that a user who likes one POI might like another POI; this likelihood is determined by the strength of the similarity between two POIs. The similarity is imprinted in the social network by its own users and might comprise a variety of components that make this information suitable for recommendation. For example, POIs may be grouped in a collection based on their categories or a characteristic of their categories (type of music for a bar, type of food for a restaurant etc), or they might signify trips with time and distance constraints. Moreover, items in the same list may exhibit complementary characteristics as this is often the desired behavior of tourists in holiday trips. As a result, two POIs with high similarity value may not necessarily have similar features but display similarities based on other criteria.

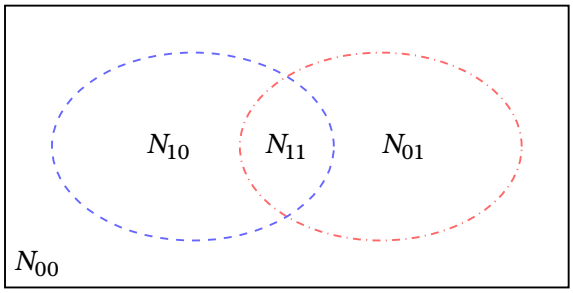
Overall, users of the social network, using their own judgement, place similar POIs in the same list as a means of grouping them under the same name or context. As a result, these POI lists can provide essential information regarding the homogeneity and diversity of the places contained within them, properties that can be exploited by a recommendation system. Ultimately, a recommendation system is only useful to the social network users so it is sensible to utilize the definition of similarity as portrayed by these same users. These similarities are the building block for the method and are used for creating the personalized recommender system.

Inferring the pairwise similarities from a given bipartite graph is often referred to as projection, an extensively used method for compressing information about bipartite networks (Zhou et al., 2007). The one-mode projection of a bipartite network $\mathcal{G} = (X, Y, E)$ onto X (X projection for short) is a weighted, complete, unipartite network $\mathcal{G}' = (X, E')$ containing only the X nodes, where the weight of the edge between i and j is determined by a weighting function $\beta_{\mathcal{G}}(X_i, X_j)$. The weighting method may not necessarily be symmetrical but, in the proposed method, a simpler approach is taken with commutative weight functions so that $\beta_{\mathcal{G}}(X_i, X_j) = \beta_{\mathcal{G}}(X_j, X_i)$, resulting in an undirected projection. Typically, the weight function expresses a form of similarity among the vertices in order to preserve the semantics of the original graph, which is perfectly suited in this scenario as it identifies the goal of the proposed method: to create the similarity matrix among the POIs.

While the projection allows to capture and quantify the similarity among the POIs that

Table 4.3: Summary of the similarity weighting methods $\beta_{\mathcal{G}}$. For two POIs x and y, the notation of this table is presented in Figure 4.6. It is noted that n'_t is the number of POIs contained within list t in order to complete the notation.

β	Name	Formula	Description
aa	Adamic	$\sum_{t \in N_{11}} \frac{1}{log(n'_t)}$	Adamic/Adar index
is	Intersection	n_{11}	Number of common lists
jac	Jaccard	$\frac{n_{11}}{n_{10}+n_{01}+n_{11}}$	Intersection over union
ka	Modified MI	Karagiannis et al. (2015)	Modified MI
ku	Kulczynski-2	$\frac{n_{11}(n_x+n_y)}{2n_x n_y}$	Intersection over harmonic mean
mi	Mutual Information	Manning et al. (2008)	Mutual Information of sets
cos	Ochiai	$\frac{n_{11}}{\sqrt{n_x n_y}}$	Cosine similarity or intersection over geometric mean
ov	Overlap	$\frac{n_{11}}{min(n_x,n_y)}$	Intersection over minimum
ρ	Phi	$\frac{n_{11}n_{00}-n_{10}n_{01}}{\sqrt{n_x*(n-n_x)*(n-n_y)*n_y}}$	Pearson correlation coefficient
sr	SimRank	Jeh and Widom (2002)	Iterative calculation of SimRank
f1	Sørensen	$\frac{2n_{11}}{n_x+n_y}$	Sørensen–Dice index or F1 score or intersection over arithmetic mean



$n_{11} = |N_{11}|$ – number of common lists
$n_{00} = |N_{00}|$ – number of lists in neither POI
$n_{10} = |N_{10}|$ – number of lists only in P_1
$n_{01} = |N_{01}|$ – number of lists only in P_2
$n_x = n_{10} + n_{11}$ – number of lists in P_1
$n_y = n_{01} + n_{11}$ – number of lists in P_2

Figure 4.6: Demonstration of the set theoretic terminology for the definition of the projection functions. Two POIs x (blue dashed) and y (red dashdotted) are shown as sets of the lists in which they are contained.

is imprinted by the social network users inside the venue lists feature of Foursquare, it is less informative than the original bipartite graph and, thus, an appropriate weighting method $\beta_{\mathcal{G}}$ is required, that minimizes this information loss. There exists, however, no universally accepted weighting method of minimizing information loss and, as a result, a selection of a selection of set and graph theoretic functions is utilized that expose the similarity among the POIs, which are given in Table 4.3. The coarse functionality of these weighting methods is that two POIs with more common lists will be more likely to be related or similar.

Most of the weighting methods that are proposed are trivial to compute as they simply rely on set theoretic measurements, but others are more computationally demanding. For example, the original SimRank has large time and space requirements (Jeh & Widom, 2002). However, the process of creating the similarity matrix needs to be executed only once, or when the primitive list data need to be refreshed. Afterwards, the recommendation algorithm can be performed directly over these projections.

A projection or a similarity matrix using a weighting method will be referred to as S or S_β when it identifies a specific projection, with $S(x, y)$ denoting the similarity value between x and y.

4.4.3 User Profile

The user profile is utilized to implement the personalized nature of the recommendations in the proposed approach. The profile is a preference vector for a single user, which corresponds to a vector of POIs, each one of which has a numerical preference value attached to it. For these values, the 5-level Likert scale ranging from “Strongly Uninteresting” to “Strongly Interesting” is utilized, with one neutral level, to encode the preference magnitude. The Likert scale is, then, transformed to the necessary arithmetic scale in order for it to be usable by the recommendation system. One option is an integer scale from -2 to 2 that corresponds to the elements of the Likert scale 1-by-1. This arithmetic scale makes more sense in a semantic standpoint and will be justified in the next subsection. It is worth noting that the proposed algorithm can operate under

any number of Likert levels as long as it is symmetric and a higher value corresponds to a higher preference. In fact, it can be applied on profiles with decimal preferences as well but, since it is dependant on human input, a 5-level Likert scale is widely used and easier to be interpreted.

Naturally, POIs in the profile need to exist in the dataset with each POI belonging in at least one list, so that the similarities with other POIs can be expressed. However, most of the weighting methods cannot work reliably with just a single list, in particular the ones that rely on the intersection (n_{11}). A POI having very little representation in the lists is easier to have its intersection with other POIs trivialized to zero and, thus, most of its similarities to other POIs also being zero. As an example, a POI belonging to a single list in the dataset, may only have 2 similarity values with another POI, one zero and one non-zero on most typical set theoretic similarity measures. These values correspond to the situation of that single list being common in both POIs or the second POI not being present in that list. As a result, having a significant number of POIs with weak list representation can degenerate the dataset to such an extent where the majority of similarity values will be zero and most POIs unable to be distinguished from one another. For this reason, it is suggested that the POIs in the profile have a strong presence in the bipartite graph. In addition, the quality of the profile can be affected by the amount of POIs as well as the diversity of the preference values within it; a user profile that covers the full range of the scale and utilizes all the available levels can possibly better portray the fine differences among the POIs in the profile.

Finally, it is important to distinguish the context of the profile and the recommendation, which may differ. For example, the profile might be referring to the city of residence of the user while the recommendation context will typically be the user's trip destination. The method can be applied on this scenario too as long as it is technically feasible with the presence of lists connecting POIs among the contexts.

For a set of POIs $p_1, p_2, \dots, p_m$, the notation that is used is P, where $P(i)$ is the preference association for p_i and, for the purposes of the experimentation, is an integer scale in $[-2, 2]$ corresponding to the 5-level Likert scale.

4.4.4 Recommendation Function

The core of the approach consists of the personalized recommendation algorithm that is capable of assigning relative preference scores to arbitrary POIs based on the similarity matrix and the personal profile of the user. Given a similarity matrix S and a user profile P, for an arbitrary POI $q \notin P$ the relative preference score w predicted by the algorithm is defined to be

$$w(q) = \sum_{i \in P} P(i)S(i, q), \tag{4.1}$$

which corresponds to the weighted sum of the similarities of the profile POIs with q, and the weights being the profile preferences themselves. The concept of a weighted

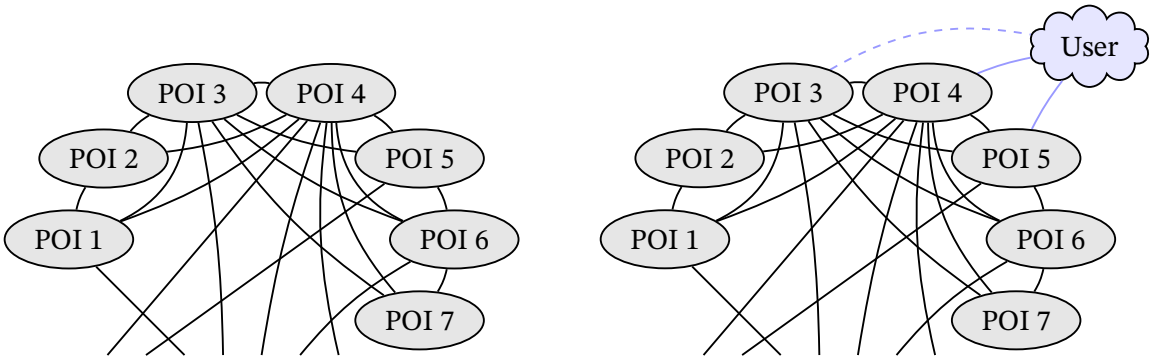
sum has also appeared in Sarwar et al. (2001), where the pairwise similarities between items and the user preferences are being utilized for recommendations.

The premise of this formula is to provide a higher relevant score w to POIs that have high similarity values with POIs in the profile that have high preference scores. For example, a POI q with high average similarity with the profile POIs ranked high (2 in the scale previously defined) will score high in the preference prediction, because of the weighted nature of the equation. Naturally, it is reasonable to assume that POIs that are very similar to "interesting" POIs will likely be interesting themselves. In essence, the algorithm driven by Equation 4.1 is based on the intuition "*Recommend the most related POIs to the POIs that the user finds interesting*".

The nature of this equation leads to interesting properties regarding its applicability and physical interpretation. First, the preference score w can be interpreted only relatively to other POIs and, hence, defined as *relative preference score*. This is due to the fact that it depends on the values of the similarity scores, which also have magnitudes with meaning only relative to others. As a result, the sign of the relative preference score is also not indicative of the item relevance to the specific user and, hence, a negative score might not necessarily imply a negative relevance. Moreover, the application of this equation across different profiles is not possible due to the absence of normalization and the scores cannot be interpreted relatively across users. Because of this, it is only meaningful to use this score when comparing two POIs and for a specific user profile, for example to answer the question "*Which of these two POIs is more relevant to the interests of this user?*". Finally, it is now clarified why a symmetric scale $[-2, 2]$ used in the profiles is more natural instead of another, such as $[1, 5]$. A symmetric scale guarantees that POIs that have been marked as uninteresting will have a negative impact; the method will be subtracting their most similar POIs while being encouraged to include their least similar ones. Similarly, neutral profile values that correspond to the arithmetic value of zero are irrelevant and can also be omitted in the symmetric scale.

Figure 4.7 shows one bipartite projection, where a weighted edge represents the magnitude of similarity between the adjacent POIs. A user and their profile can be seen in the projected graph as a new vertex with edges across the profile POIs and weights equal to the preference scores. The recommendation algorithm considers the projection as well as the profile to calculate a relative preference score for another arbitrary POI that exists in the dataset.

Based on Equation 4.1, the recommender system can be constructed by simply including the top quantities for a similarity matrix S and a user profile P. In particular, the algorithm can make preference predictions for a set of POIs in an area, rank them with respect to their predicted relative weights and offer the top-N recommendations. The recommendations can also be filtered based on categories, areas, or other user defined criteria. These calculations are simple to process and in certain cases can be performed



(a) A single abstract bipartite projection.

(b) The projection along with a user.

Figure 4.7: (a) The transformation of the bipartite graph into its projection using a similarity measure and (b) the same graph with a user attached. In this case, the user states preferences for POI 4 and POI 5 and the algorithm uses this profile along with the similarity values of the projection to assign a relative preference score to POI 3 (dashed edge).

in real time, as S is already computed once. The complexity of the recommendation is, hence, proportional to the filters assigned to the recommender system.

An inherent limitation of the recommendation algorithm is due to the relevant nature of the preference predictions. In an extreme scenario where all of the candidate POIs in the recommendation context are irrelevant to the user and their interests (i.e., they do not like any of them), the recommendation algorithm will result in personalized suggestions that are also irrelevant. The reason is that a high preference prediction does not necessarily mean an interesting POI but rather a POI that is more interesting than the rest of the POIs in the context. This restriction is a common aspect of ranking algorithms and, typically, can be circumvented by expanding the recommendation context in such extreme cases.

As a closing remark, it is worth considering the data space of both the input of the recommender algorithm and its output. In particular, the input is a set of POIs along with their respective preference weights (profile) and its output is a ranking of POIs denoting the predicted preference ranking of that particular user. It can be seen that the possible states of the input is exponential with respect to the number of POIs in the profile. For example, if there is only 1 profile POI, the possible states are 2: a positive preference weight or a negative one; a zero preference score makes the recommendation infeasible and is ignored. In general, the maximum possible states of the input are 5^m, assuming the 5-level Likert scale is used and the POIs in the profile are m. In contrast, the output states are increasing with factorial order with respect to the number of POIs in the context, which are in most cases more than the POIs in the profile. However, given that the recommendation method is deterministic, the number of recommendation outcomes are bound by the input states, which grow much slower than the maximum output states. As a result, it is a good practice to include the most amount of POIs in the profile that are possible in order to increase the possible outcomes

of the algorithm.

4.5 Experimental Results

The effectiveness of the recommendation methodology is evaluated against the case study dataset described in Section 4.3. In particular, the two aforementioned areas *Thessaloniki* and *Kassandra* are used and the method is evaluated using POIs and POI lists in these regions. The evaluation is threefold and consists of the *preference evaluation scheme* (Section 4.5.1), the *recommendation evaluation scheme* (Section 4.5.2) and the *offline evaluation* (Section 4.5.3). During the preference evaluation, the relative preference predictions of Equation 4.1 are assessed on arbitrary POIs, which can have positive or negative impression on the experiment users. For the recommendation evaluation, the same method that was used in the TREC 2016 Contextual Suggestion Track (Hashemi et al., 2016) is utilized, which evaluates the system as a recommender, only includes top preferences for each user and is performed in two stages. Lastly, for the offline experiments, the system is cross validated by using some of the POI lists as virtual profiles in order to complement the online survey responses. During all evaluation schemes, the effectiveness of the different similarity measures is also assessed and observations are made about the properties of each one. As for the ground truth, a user study in the form of a questionnaire was incorporated, with participants familiar with the attractions of the user case areas.

The implementation of most of the similarity measures on the dataset can be done by simply treating the POIs as the set of lists within which they are contained and applying the formulas in Table 4.3. There are only two special cases of measures that are worth mentioning in an implementation perspective: the Adamic/Adar index and SimRank. Regarding the Adamic/Adar index, the quantity n'_t (number of POIs within list t) is considered to be the number of POIs that the Foursquare API advertises as belonging to list t, which is the intended way of using this index. This quantity might not be in agreement with the number of POIs in list t in the dataset due to the way that the data were acquired, which was biased towards the POIs, and is, therefore, smaller (or equal) than the true n'_t. For the implementation of SimRank, the dataset is treated as a bipartite graph as shown in Figure 4.5b. For the implementation of SimRank, the original computation algorithm is used. In addition, SimRank has a parameter c which is usually set to 0.8 (C. Li et al., 2010), but the value 0.6 is used too as a means of relative comparison and they are denoted as *sr8* and *sr6* respectively.

4.5.1 Preference Evaluation

The preference evaluation scheme aims to assess the effectiveness of the relative preference scores that are returned by Equation 4.1, which is the basis of our algorithm. Simultaneously, we are doing a comparative analysis of the results and performance of

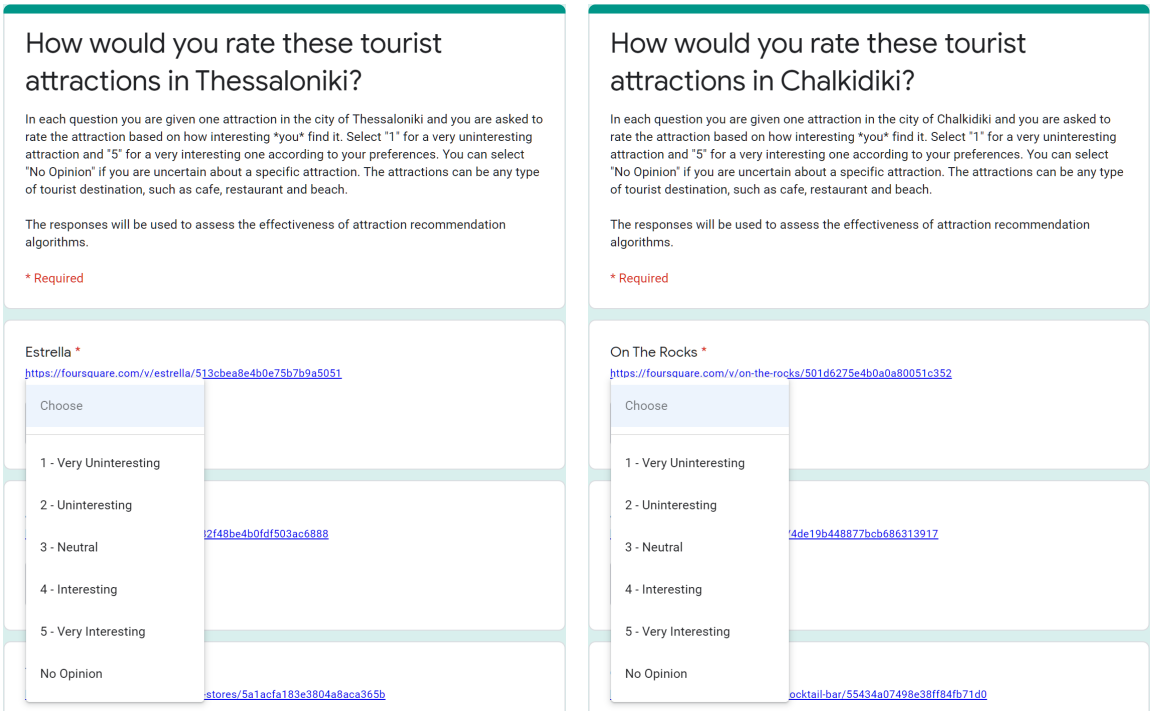


Figure 4.8: Screenshots of the user survey in the two regions of Thessaloniki and Kassandra.

all projection weighting methods.

Initially, we select the POIs that will participate in this experiment based on several criteria and conclude with 19 POIs in Thessaloniki and 11 POIs in Kassandra. These POIs are attractions that match the tourism type of these areas, namely cafes, bars, beaches or restaurants. They are selected with Foursquare rating uniformity in mind in an attempt to collect answers that contain a more balanced preference distribution. The ratings are ranging from 6 to 10, as there were very few to no venues with rating less than 6. For POIs that were around the same rating, we intentionally considered the ones with the most amount of lists so that the users participating in the survey were more likely to have visited them or to have a developed opinion about those. We refer to these 19 and 11 POIs as *survey POIs*.

The survey participants were asked to rate each of these POIs based on how interesting they personally find them. The survey options were compatible with the 5-level Likert scale and were:

1. Very Uninteresting
2. Uninteresting
3. Neutral
4. Interesting
5. Very Interesting

We also included "No Opinion" to account for cases where users wanted to refrain from expressing an opinion and ignore these responses in our analysis. Overall, we received 31 user responses for Thessaloniki and 16 responses for Kassandra but we only considered the responses that had at least half of the POIs answered, which were 28 and 14 respectively. Cropped screenshots of the survey for both areas are displayed in Figure 4.8.

The survey POIs and the survey responses are utilized both as profile and ground truth via the leave-one-out cross-validation method. According to this method, one POI k is held out from the profile and its rate is attempted to be predicted; the process can be summarized in the following steps:

1. Compute the relative recommendation score $w(k)$.
2. Normalize $w(k)$.
3. Compare the normalized w vector against the ground truth vector.
4. Repeat the process over all users and projections.

For each missing POI k, the relative preference weight $w(k)$ is computed via Formula 4.1 with k itself excluded from the user profile. Then, we apply a small transformation to $w(k)$, specific to this experiment, due to technical limitations of the cross-validation method. In particular, because each profile with an excluded POI is considered a different profile, the relative scores $w(k)$ cannot be interpreted relatively across these profiles, even if they refer to the same user; more information has been established in Section 4.4.4.

As a simple way to overcome this technical difficulty, we normalize the predicted scores $w(k)$ with the sum of the profile preferences $\sum_{i \in P} P(i)$. Hence, we get a vector of predicted preference values, one for each survey POI which is then compared to the ground truth vector of the preferences stated in the user profile. The comparison is done using the Pearson correlation coefficient, which is a measure of the linear association between these two vectors, and Kendall tau-b correlation (Agresti, 2010). Tau-b correlation is a rank correlation measure and a generalization of the Kendall tau-a coefficient that accounts for ties in the input lists, specifically present in the distinct 5-level preferences of the survey. This process is repeated over all users and projections.

Figures 4.9 and 4.10 display the evaluation results of Thessaloniki and Kassandra respectively as box plots. The results are grouped by similarity measure to better convey the differences among them. It should be noted that, regarding the rank correlation, there exists a maximum τ value that the algorithm can achieve due to the discrete survey answers, where two POIs can be tied on the same rank. The same situation is very unlikely to occur in the output of the algorithm due to the floating point nature of the operation, which in practice disallows ties on the same rank. As a result, the maximum

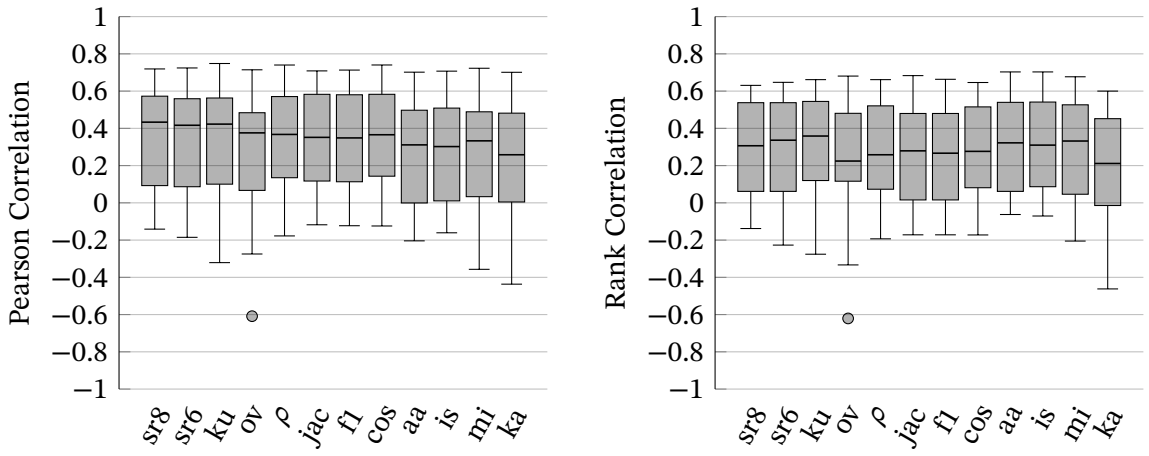


Figure 4.9: Preference evaluation scheme results for Thessaloniki (28 data points). Each box displays the 4 quartiles of the user distribution. The measures are ordered by the median of the Pearson correlation without considering the outliers.

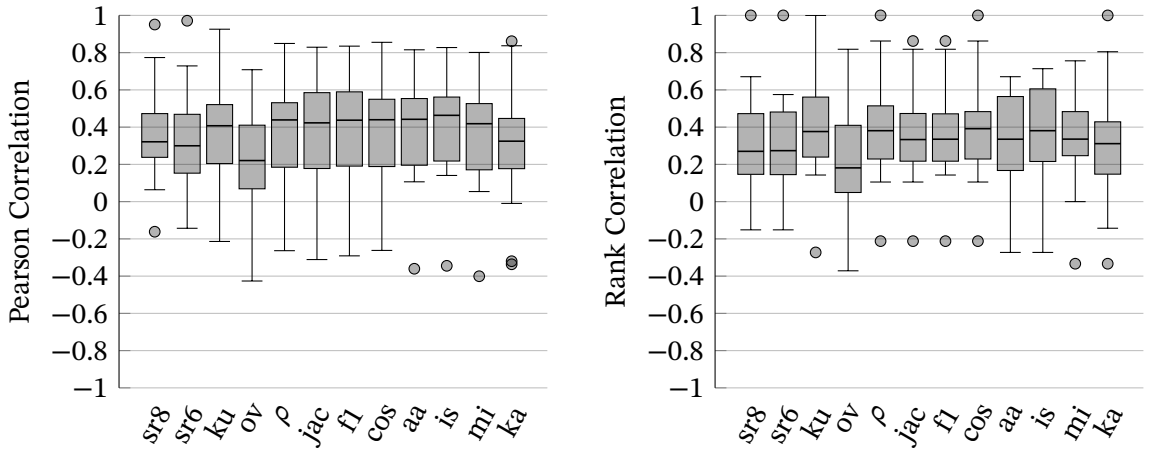


Figure 4.10: Preference evaluation scheme results for Kassandra (14 data points). Each box displays the 4 quartiles of the user distribution. The ordering is the same as Figure 4.9 for consistency.

correlation can be shown to be equal to the correlation of the ground truth vector with the "best" possible untied vector (the flattened ground truth vector without ties):

$$\tau_{max} = \frac{T(n) - \sum_i T(t_i)}{\sqrt{T(n)}\sqrt{T(n) - \sum_i T(t_i)}}, \text{ where } T(x) = \frac{x(x-1)}{2} \tag{4.2}$$

and $t = [t_1, t_2, t_3, t_4, t_5]$, with t_1 being the number of POIs in the "Very Uninteresting" group, t_2 being the number of POIs in the "Uninteresting" group etc. For this reason, we only show the ratio of tau-b correlation to τ_{max} on the rank correlation plots. Following Equation 4.2, it is also evident that the quantity τ_{max} may be different for each profile. Increasing the number of possible ranks in the survey would diminish this behavior but at the cost of user convenience.

In general, we have mixed feelings about the preference evaluation as there exist both strong and weaker results. In particular, it seems as the values are consistent, with most of the similarity measures firmly around the same restricted range of values. Interestingly, there is a small discrepancy between the Pearson and rank correlation as there does not appear to be a clear winner in both settings. Furthermore, in all settings, the majority of the projections outperformed the modified MI index (ka) and the overlap coefficient (ov), the latter of which also contains the only outlier in Thessaloniki. Both versions of SimRank (with parameters 0.6 and 0.8) are the top performing similarity measures in Thessaloniki and their top two quartiles can be considered as good results (Akoglu, 2018). Although this statement is not true for Kassandra too, this, along with the observation that results in Kassandra appear to be better, can be attributed to the smaller sample size and, therefore, the reduced difficulty of the problem. As a result, the two geographic regions cannot be directly compared to one another. We argue that, even with this small size of questionees set, our complex method and its results indicate firm evidence of the fact that lists contain information useful in the context of POI recommendation and prompts us to perform experiments on our system as a recommender.

4.5.2 Recommendation Evaluation

The recommendation evaluation scheme aims to assess the effectiveness of the personalized recommendations that result from assuming the top quantities of a user profile and projection scenario. The major difference between this scheme and the preference evaluation is that now we only assess the POIs that are most relevant to the users: those with the top relative preference scores. We are using the top-5 quantities as this appears to be the most dominant setting. Similarly, as in the preference evaluation scheme, we are also doing comparative analysis among the weighting methods.

The process is based on the TREC 2016 Contextual Suggestion Track (Hashemi et al., 2016), where the responses of the Section 4.5.1 survey are used exclusively as profiles.

Specifically, by using each response as profile, we produced the top-5 results for all weighting methods and aggregated them. While some of the recommended POIs were common across the weighting methods, the cardinalities of the recommendations' unions were all more than 5 POIs. Since every user rated the profile POIs differently, the resulting aggregated POIs that contain the combined recommendations of all the weighting methods are different for each rater. The combined recommendations of a rater are then evaluated against another survey that is forwarded to that rater. This new questionnaire has the same structure as the previous survey in terms of the Likert scale levels and is personalized for each user as the recommendations are different for them. The recommendations are performed for the same context as the profile so that the same user that receives the additional survey is more likely to have prior experience about the recommended POIs. The recommendation evaluation scheme was only done in Thessaloniki as we did not manage to get enough responses from users to complete the experiment in Kassandra. Specifically, for Thessaloniki, we received an extra 8 responses, of which 1 had very little information filled in and, thus, we consider 7 of these replies.

The recommendation system is evaluated using the following measures:

1. Precision at rank 5 (P@5)
2. Normalized discounted cumulative gain at rank 5 (NDCG@5)
3. Mean reciprocal rank (MRR)

which were also part of the evaluation of the submissions of the TREC 2016 competition. P@5 shows the fraction of the personalized suggestions that the user liked, where "liked" refers to the evaluation as "Interesting" or "Very Interesting"; the other 3 Likert options were considered as non relevant. Since the suggestions were 5 for each weighting method, the values of P@5 can be 0, 1/5, 2/5, 3/5, 4/5 or 1, with 1 meaning that the user liked all 5 of the algorithm recommendations and is considered the perfect score. However, because there were responses marked as "No Opinion", in these cases we use P@4 or even P@3, depending on how many answers were missing. Moreover, the NDCG@5 measure is an indication of the ranking quality of the 5 personalized recommendations and, like DCG, does not take into account whether the recommendation was relevant or not but only the ranking of the top POIs. For this, we are using 5 ordered, arithmetic classes, one for each response option. A value of 1 corresponds to the scenario where the user evaluated the 5 suggestions in the same ranking as the algorithm did. Finally, the MRR measure shows the average of the ranks of the first relevant ("Interesting" or "Very Interesting") suggestions. A value of 1 means that on all cases the first recommended POI was liked by the evaluator. A value of 0.5 means that, on average, the first recommended POI was irrelevant but the second was relevant. Lastly, a value of 0 means that there was no relevant POI in the recommendation list. The nature of MRR dictates special considerations because of the presence of "No Opinion" values in the surveys: where the MRR cannot be accurately calculated, we

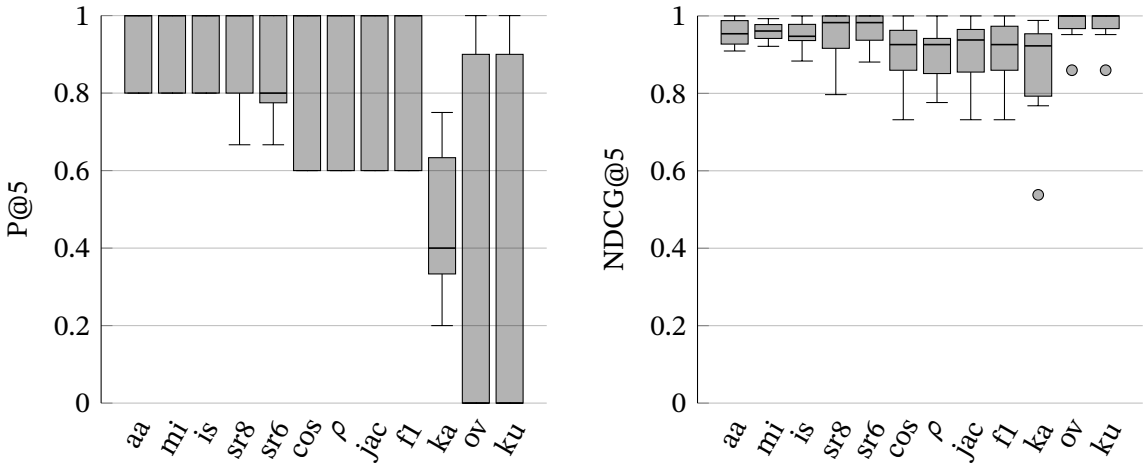


Figure 4.11: Recommendation evaluation scheme results for Thessaloniki (7 data points). Each box displays the 4 quartiles of the user distribution while the average is displayed as a diamond symbol. The measures are ordered by the average P@5.

Table 4.4: Recommendation evaluation using the MRR measure. The ordering is the same as Figure 4.11 for consistency.

	aa	mi	is	sr8	sr6	cos	ρ	jac	f1	ka	ov	ku
MRR	1.00	1.00	1.00	0.88	0.80	0.93	0.79	0.83	0.83	0.72	0.33	0.67

do not consider it at all in the average calculation. One example would be a ranking where the first relevant POI was below one or more POIs marked with “No Opinion”, in which case it is unclear what the value of MRR should be. As a result, there were some weighting methods with just 4 user responses considered; this is the also the minimum amount.

Figure 4.11 presents a box plot of the recommendation evaluation with P@5 and NDCG@5. The Adamic/Adar index (*aa*), MI (*mi*), intersection (*is*) and SimRank (*sr8*, *sr6*) appear to perform almost flawlessly, with average precision over 0.9; in practice more than 9 out of 10 POIs on average were correctly recommended. However, because our sample size is relatively small, it is unclear which of these is the best. Interestingly, most of the similarity functions performed very well at NDCG@5, even the weaker measures and a possible explanation is due to the granularity of the method, where only 5 discrete values are present, of which the two highest were used almost exclusively. Table 4.4 shows the MRR values of the experiments, which seem to be in agreement with the P@5 results. Deviations from this behavior is because MRR only considers the first relevant POI and, as a result, is biased towards mostly the top result, instead of P@5 that assigns equal weight to all 5 of the suggestions.

Further examination of the results reveals an interesting observation about the varied performance of SimRank. The only POIs that both versions (0.6 and 0.8) of this weight-

ing method failed with respect to the P@5 measure appear to be 3 distinct recommended POIs within a single list each and no other metadata, such as rating, reviews or other information. It is possible that these POIs are wrong submissions that happened by some Foursquare users in an effort to identify a check-in or another spatially identified action. In fact, 2 of these POIs, 7 months after the acquisition of the dataset, are now not part of any list in Foursquare, possibly because it was an error that was corrected by the users themselves. We believe that these POIs should not have been included in the experiment in the first place because of the trivialized information contained in their records and an extension to this experiment would confirm the increased accuracy in this scenario. Even though we did not perform any filtering in the dataset besides the necessary POIs that are contained within at least one list, we still consider the existing results very strong in terms of recommendation effectiveness.

4.5.3 Offline Evaluation

Finally, the offline evaluation aims to complement the limited survey responses of the online experiments by utilizing user generated information that is already present in the LBSN. In this section, we describe the offline experiment via which we evaluate our recommendation method by considering the POI lists as virtual user profiles. Despite the lists themselves being part of the dataset, the assumption that lists can be used as user profiles is reasonable since they are being created by users to represent one dimension of their own interests. For this study, we have adopted evaluation techniques that are common in the domains of session-based recommendation or next-track (playlist) recommendation that is further explored in the related work section. In particular, a list is being considered as a user, and the recommender is used to predict held-out elements of the list. This evaluation process is explained in more detail below. Lastly, in this experiment, we have included 3 popularity baselines about the POIs: the number of likes as reported by Foursquare, the rating of the POI as obtained by the Foursquare API (decimal number in the scale $[0, 10]$) and the degree of the POI (number of lists containing that POI).

When considering a certain list as a virtual user profile, that list must be removed from the dataset before the recommendation process. Since a removal from the dataset can interfere with the pairwise similarity factors, the complete bipartite projection must also be computed every time we use a list as a profile. As a result, the lists that contain POIs with singular degree cannot be used as profiles as those POIs would be missing from the dataset and, despite existing in the profile, would not be eligible recommendation targets. For consistency with the online experiments we also exclude the lists that contain POIs with degree 2 as their removal from the dataset would leave these POIs existing in a single list. Furthermore, we did not consider lists with less than 6 POIs as profiles as an attempt to increase the list degrees; for comparison, the online surveys consisted of 11 and 19 POIs for Chalkidiki and Thessaloniki respectively. Lastly, we only utilized the lists with the *collaborative* field set to *false* as profiles. This restriction

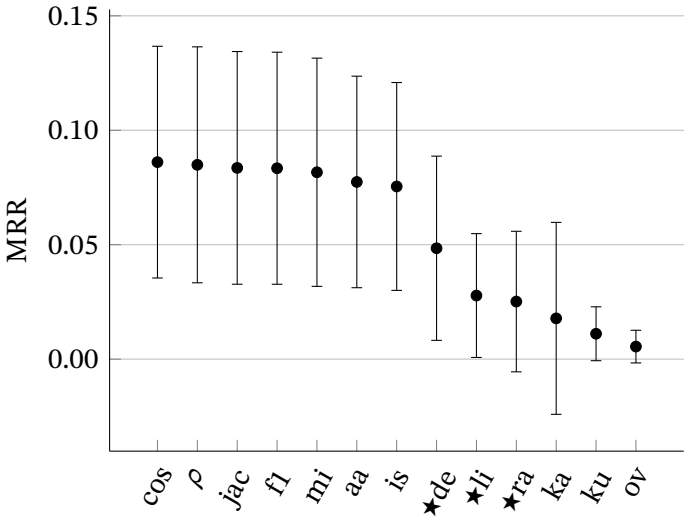


Figure 4.12: Results of the offline evaluation. The error bars display the average MRR and the standard deviation. The baseline measures degree (*de*), likes (*li*) and rating (*ra*) are marked with a star (★). Each projection corresponds to 747 virtual profiles.

constitutes an additional check to further establish that only the lists that were made by a single person are being used as profiles, since collaborative lists might correspond to inconsistent interests. The lists with this property were less than 2% of the total lists. Given the aforementioned constraints, there are 747 lists that do not fall into one of the filtering categories and are used as profiles in this experiment. Despite using only these 747 lists as profiles for practical reasons, all of the lists are being utilized as part of the dataset.

For each list being used as a virtual profile, we perform a leave-one-out cross-validation and attempt to predict the missing item from the list in a way similar to the preference evaluation scheme. In particular, for every missing item, the rest of the profile is used to create a ranked recommendation list for all POIs in the dataset based on their relative preference weights (Equation 4.1) and note the position at which the correct (missing) one appeared. Because lists do not explicitly contain preferences as they were not created to represent profiles, we consider all POIs in a list to have the best preference (Very Interesting). The MRR measure is used to capture the average rank at which the missing item was predicted. The MRR value of 1 corresponds to the situation where all POIs of the virtual profile were correctly predicted in the first position of the ranked recommendation list, a value of 0.5 means that on average they were on the second rank.

Figure 4.12 shows the results of the offline evaluation grouped by similarity measure. Each projection is identified by the 747 lists that we use as virtual profiles and display the average MRR and the standard deviation of the MRR as an error bar. In this scenario, the average MRR corresponds to the average of the average reciprocal rank. The plot includes the baselines of degree, likes and rating marked with a star symbol (★) as mentioned previously. For these baselines, the recommendation ignored the profile and returned the POIs sorted based exclusively on the respective measure. Essentially, the average of the average rank of the missing POI in a list was 17.75 for Ochiai (*cos*),

measured under 3 397 POIs, which corresponds to 0.5 percentile average position. Another observation is the performance of the ka, ku and ov measures which are below baseline, a property that is in alignment with Figure 4.11. The performance of the other measures (*cos*, ρ, *jac*, *f1*, *mi*, *aa*, *is*) is also in alignment with the online experiments and their effectiveness is almost identical. As a result, there does not appear to be a similarity measure with clear advantages over the others in this group, an observation that is consistent with the online recommendation experiment. Moreover, these 7 similarity measures are considerably more effective than the baselines and produce better recommendations than simply recommending the most popular. Finally, among the baseline recommenders, degree appears to be the most prominent while likes and rating have similar effectiveness.

Due to technical restrictions and its computational complexity restrictions SimRank is not included in this experiment. Specifically, every projection had to be re-created for each of the 747 virtual profiles because it had to be removed from the dataset in advance. As a result, this process was prohibitive for SimRank as the similarities had to be recomputed for each virtual profile. However, based on the results of the online study and the consistencies with the offline experiments, it is possible that SimRank would also perform equally well with the rest of the 7 top measures.

4.5.4 Results Discussion

Our twofold evaluation analysis yields interesting combined results about the algorithm and the similarity measures that we use, and allows us the juxtapose the results and compare our approach on two settings. Initially, we discuss the observation that our approach appears to be better in the online and offline recommendation evaluation but only moderate in the preference evaluation. This can be due to a combination of reasons, for example the algorithm may simply be more effective at identifying the top quantities, or the top POIs contain more information (such as more lists) in the LBSN and, thus, easier to extract. Nevertheless, neither of these approaches is trivial as it has to scale for a large set of POIs in the recommendation context, possibly in the magnitude of thousands or more.

Regarding the use of the similarity measures, it can be observed that some of them are effective when measured via one evaluation method but mediocre via the other scheme. This may indicate that not all similarity functions are appropriate for every circumstance and each one may be used more effectively depending on the individual problem. In addition, the results suggest that not all similarity functions are appropriate for the study of this chapter. For example, the overlap function (*ov*) appears to deliver weaker results in comparison with the rest of the similarity measures.

The combination of the online evaluation schemes, however, seems to converge into the conclusion that SimRank is marginally more consistent and effective. Specifically, SimRank with parameter 0.8 (*sr8*) is better than the 0.6 version (*sr6*), but this difference

is negligible. Interestingly, SimRank is the only global similarity index that takes the properties of the whole network into consideration, as opposed to the local similarities that only consider local information (Schall, 2015, Section 2.2). While SimRank leverages the whole bipartite graph to compute the similarity among x and y, the rest of the measures only take into consideration the lists of x and y. An exception to this is the Adamic/Adar index which is a unique measure that considers local information (the intersection of the lists between two POIs) as well as the degrees of the intersected lists as an additional step. Thus, if the set theoretic measures are considered as 1-step operations (the adjacent nodes of the POIs), the Adamic/Adar index is a 2-step operation (the adjacent of the adjacent of the POIs) and the global indices are a multi-step operation (values that propagate through the graph until convergence). The same concept is also described in Goyal and Ferrara (2018) using the terms *first-order*, *second-order* and *higher-order proximity*.

Based on these observations, the pattern that emerges is that global similarity measures perform better than local similarity functions in these POI list graph based recommendations. Specifically, while SimRank appears to be on average a well balanced projection in terms of effectiveness, the Adamic/Adar index, which is a semi-global index, also seems to be on par, particularly on the recommendation evaluation scheme. Whether this hypothesis is true, and to what extent, remains to be seen in future work experiments and in a larger scale.

The instances that our approach failed can be attributed to the difficulty of determining when a POI is similar to another. Each individual may have a different judgment when asked to decide if two POIs are similar as they are subjective to their own preferences and, hence, failures will naturally occur in such recommendation systems. For example, when comparing two cafes, some may conclude that they are similar based on the kind of music that they play while others based on the art style or atmosphere of the two POIs. As a result, similarity, in a general sense, may be difficult to express using a single measure for every occasion. In fact, since lists are user-generated content, they might contain a form of average of the criteria of what people think constitutes similarity, and, thus, one may have to selectively pick lists that express a particular form of similarity to satisfy perspectives of different users.

On a related note about the difficulty of determining the similarity between two POIs, it is worth mentioning how Foursquare defines this similarity. In an article published by the Foursquare engineering team[4], Foursquare defines a “similar place” as a “venue similar to the specified venue” and is calculated using an aggregation of various metrics:

- Covisitation: The premise of using covisitation to calculate venue similarity is that if a lot of people tend to frequent two venues, they may be similar to each other. Covisitation calculation utilizes cosine similarity as the scoring mechanism.

[4]https://medium.com/foursquare-direct/finding-similar-venues-in-foursquare-cf535d9028ee

- Category: Foursquare uses maximum likelihood estimates to determine the probability for a venue labeled with categories x and at least one other to be also labeled with category y. This measure is used to establish a similarity score between different categories with the same root (e.g. Dim Sum restaurant and Hong Kong restaurant) and consequently venues falling under these.
- Tastes: Tastes are user-contributed tags associated with venues and can include specific dishes (like lasagna) or certain atmospheres (cozy, romantic) or any other trait of a venue. Taste similarity is calculated by looking at the tf–idf for matching tastes in two venues.

4.6 Discussion: Recommendation Diversity

Studying the accuracy of recommendation systems is critical for the assessment of the recommendation suggestions for the users or other interested parties. There exist, however, other perspectives of the effectiveness of recommender systems other than their accuracy. In particular, Kaminskas and Bridge (2017) present and discuss more elaborate objectives than accuracy of recommendation systems based on collaborative filtering and specifically focus on the decrease of popularity bias and the increase of diversity, novelty and serendipity of a recommended list of items. The paper also consists a review of the existing literature regarding these objectives and present definitions, comments and algorithmic approaches.

In this section, we focus on concept of *diversity*, which is an interesting idea that has gained recent attention in recommendation systems. The recommendation diversity typically measures the average pairwise distance among items in the recommendation list and has been suggested by Smyth and McClave (2001). More formally, if the recommended items are denoted as $R = \{r_1, r_2, \dots, r_k\}$, then the recommendation diversity is defined as

$$\text{diversity}(R) = \frac{\sum_{i \in R} \sum_{j \in R \setminus \{i\}} \text{dist}(i, j)}{|R|(|R| - 1)}, \tag{4.3}$$

where $\text{dist}(i, j)$ is a measure of distance or dissimilarity between items i and j. Often, the dissimilarity measure is directly related to the similarity measure used in the recommendation procedure but it can also be independently selected. Summarily, Formula 4.3 represents the average pairwise dissimilarity among the items in the recommendation list.

In this discussion, we present a preliminary analysis of the properties of our recommendation system surrounding its diversity. In particular, we perform independent analyses for each of the projections to uncover their underlying features related to the diversity of the respective recommender. For consistency, we use the same similarity measure that was used in the projection as the dist function in Equation 4.3 and change the semantics

of diversity as inverse diversity or *agreement*. Hence, the equation becomes

$$\text{agreement}(R) = \frac{\sum_{i \in R} \sum_{j \in R \setminus \{i\}} S(i,j)}{|R|(|R|-1)}, \tag{4.4}$$

where $S(i,j)$ corresponds to the similarity between POIs i and j as shown in Section 4.4.2. The transformation is necessary as some similarity measures used in this work may not have obvious distance counterparts, for example the unbounded value of the Adamic/Adar index. In contrast, for other bounded measures, for example in $[0,1]$, the distance is usually interpreted as the complementary of the similarity $\text{dist}(i,j) = 1 - S(i,j)$. It is worth noting that in the following analysis we consider the top-5 recommendations, i.e. $R = \{r_1, r_2, r_3, r_4, r_5\}$.

Since the same similarity measure of the projection is used for the assessment of the level of diversity in the recommendation, it follows that the quantities resulting from the direct application of Equation 4.4 cannot be relatively interpreted among the different projections as the values of the similarities cannot be compared to one another. For this reason, we discuss two indices of the recommendation diversity that stem from Equation 4.4: its *z-score* and its *p-value*. These indices normalize the diversity of each projection and makes it possible to compare and juxtapose the different projections with respect to their diversity properties, similar to the comparison performed in Section 4.5 in terms of the recommendation accuracy.

In particular, the z-score of the diversity is defined as the standard score of the agreement with respect to the distribution of all pairwise similarities in the dataset and shows how many units of standard deviation is the average recommendation similarity agreement(R) greater than the average POI similarity. The greater the z-score, the more similar are the POIs inside the recommendation list with respect to the average POI similarity and the less diverse the recommendation list. A value of 0 (zero) indicates the absence of correlation and an average similarity of the recommendations that coincide with the mean of the dataset POI similarity. Similarly, the p-value of the diversity is the percentage of pairwise similarities in the dataset that are greater or equal than agreement(R). A higher p-value indicates a higher level of diversity while a value equal to 0.5 indicates that the average recommendation similarity coincides with the median of the dataset pairwise similarities.

Figure 4.13 displays the z-score (top plot) and the p-value (bottom plot) of the diversity as a bar chart. The projections here are ordered the same way as Figure 4.11, i.e. decreasing with respect to the recommendation accuracy. It can be immediately seen that the z-scores of all similarity measures are positive, confirming the intuition that POIs similar to the profile of a user should be similar among one another. Similarly, the p-values of the measures are smaller than 0.5 indicating an related behavior. Another preliminary observation is the relation between accuracy and diversity which appears to be inverse; often, projections with high accuracy have low levels of diversity (or high levels of agreement) in the respective recommendation lists. The lowest diversity scores are

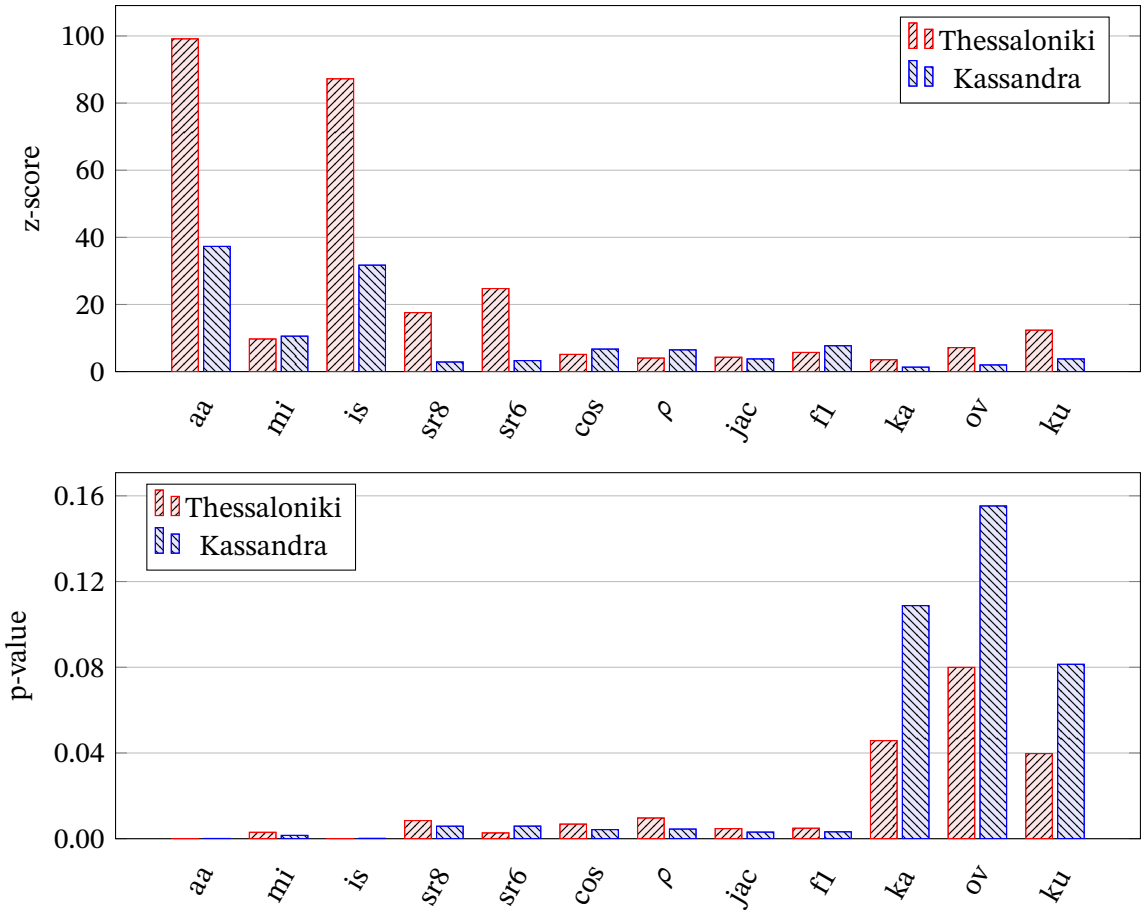


Figure 4.13: Illustration of the diversity of the top-5 recommended POIs. The z-score (top plot) increases as the level of diversity in the recommendation list declines. The p-value (bottom plot) increases with the level of diversity in the recommendation list. The order of the projection measures follows Figure 4.11 as decreasing with respect to the recommendation accuracy. A preliminary observation is the inverse correlation of the recommendation accuracy with the recommendation diversity.

those of the Adamic/Adar projection (*aa*) and the intersection (*is*), which coincidentally are the only unbounded value projections in our experiments.

4.7 Discussion: Lists

In this section we share some preliminary but interesting findings about the types of information that is inherent in the POI lists and some insights about our similarity measures and their possible use in certain applications. Specifically, in the first experiment, we attempt to identify what components of information are present in the Foursquare lists and how these are related to the criteria via which users group POIs in lists. In the second experiment of this section, we perform pairwise correlations among the different similarity measures and investigate possible relationships based on the observations. Both experiments are performed in the same dataset and specifically in Kassandra, as the high number of POIs in Thessaloniki would hinder the applicability of several of these experiments.

4.7.1 Information components in lists

The motivation of this experiment is provided by the examination and careful inspection of the list names. A common pattern that arises can be summarized in names with the common theme "Best nightlife in Chalkidiki" or similar variations. All of the words in such lists can be considered as cues for the information components in them. For example, *best* implies the existence of high rated places, which may be subjective to the preferences of the creator of that list, while *nightlife* indicates the types of POIs or their categories. Finally, *Chalkidiki* is a spatial specification of the list and it is very common among our data; from the 50 most popular lists in our dataset (with respect to the number of followers), 32 have at least one term that corresponds to a geographic region.

These observations prompt us to perform experiments and confirm them in an extended scale. These experiments allow us to make some preliminary claims as to what types of information is inherent in Foursquare lists as well as to define the applications that would benefit from these components individually. Our goal is to correlate the similarity between two POIs x and y with measures that convey the three components mentioned earlier:

1. *Rating* in the form of the Foursquare rating difference between x and y. Our hypothesis is that higher similarity value results in smaller rating difference or, equivalently, POIs within the same rating bracket should be included in the same list.

2. *Categories* as the overlap of the categories of x and y. We are using this measure instead of a direct comparison between the category of x and y because a POI

Table 4.5: Minimum and maximum Pearson correlations of projections with geographic distance, categories overlap and rating difference.

Geographic Distance	Categories Overlap	Rating Difference
0.11 - 0.20	0.14 - 0.30	0.00 - 0.33

in Foursquare can have multiple categories. For the POIs with categories $\{c_x\}$ and $\{c_y\}$, overlap is being defined as the number of shared categories over the minimum number of categories $|c_x \cap c_y|/min(|c_x|, |c_y|)$.

3. *Geolocation* in the form of the geodesic distance between x and y. Smaller geographic distance should, on average, display correlation with higher similarity.

We compute these correlations for every pair of POIs in the dataset of Kassandra and display our results in Table 4.5. For each of the three components, we aggregate all pairs of a similarity measure into their average and present the minimum and the maximum Pearson correlation coefficient out of all the similarity functions. The results of this experiment are, to a certain extent, in alignment with our observations. First, regarding the geographic distance it can be seen that the distance between two POIs is, on average, positively correlated with their similarity. The table displays the absolute value of this correlation as the magnitude was negative; the shorter the distance between the POIs, the larger the correlation. Although, the value might seem insignificant, it should be mentioned that the amount of values that are being considered are the number of pairs of POIs, which is in the order of 526^2 and even the minimum value (0.11) results in a p-value significance of less than 10^{-5}. In addition, we acknowledge that the similarity does not correlate exclusively with the geographic distance and, hence, higher correlation would not allow room for other components. Furthermore, the similarities also appear to be correlated with the overlap of the categories among the POIs, even higher than the geographic distance. Finally, a rather interesting result is the third component of information, namely the correlation with the rating difference among the POIs that indicates that POIs with similar Foursquare rating are more likely to have higher similarity values and, thus, be contained in the same list. This phenomenon may be reinforced by the fact that POIs that are already popular are more likely to be placed together in new lists. This experiment, however, requires further examination as the variance of the correlation values among the projections was very high, with some similarity measures scoring as low as zero, implying no correlation at all. A possible explanation is that users have largely different judgments of the definition of rating and it is reasonable that, as the only subjective measure in these experiments, the variance among the users is high. The other measures (geographic distance and categories overlap) are more objective that do not rely on users' preferences or judgement.

We acknowledge that there are other components that the similarities are correlated with, for example the type of music of a cafe, the art style of a restaurant and many

more. Performing these experiments on specific areas individually and juxtaposing with a larger area might also provide other kinds of information, for example what type of tourism is more prevalent in each area, which could be cultural, leisure, or other forms of tourism. The options and combinations of components that are encoded in the lists are potentially limitless and, as a result, we do not consider these experiments to be complete or concluding.

Despite this, the results indicate the presence of multiple components of information within the lists, knowledge that is acquired through the contribution of social network users' experience. The POI bipartite graph structure can be thought as a balanced combination of these multiple primitive components of information portraying the relationships and similarities among POIs, which in this work we leverage to propose this POI recommendation method.

4.7.2 Similarity measures relationships with the list components

The results of minimum and maximum correlation values of Table 4.5 leads to further interesting observations regarding the effectiveness of the similarity measures at portraying individual list components. The observations refer to the similarity measures that achieve the maximum values in that table, which can provide indications as to which measures are more appropriate for portraying individual list components. Specifically, the geographic distance measure appears to be mostly prevalent in the MI and the modified MI (*mi*, *ka*) projections with 0.18 and 0.20 correlation respectively while SimRank and the Jaccard index (*sr6*, *sr8*, *jac*) are driving the maximum value of the categories overlap correlation, occupying the 0.29 – 0.30 range. Finally, the MI and modified MI similarities also appear highly correlated with the rating difference with 0.12 and 0.33 respectively. These observations motivate us to study the quantifiable relationships among the similarity measures and possibly identify groups of projections that are more related than others with respect to their physical interpretation.

4.7.3 Relationships among similarity measures

This experiment is designed to measure the correlation among all pairs of projections by considering all the pairwise POI similarities for each projection. The magnitude of this correlation is shown in Figure 4.14 as color coded background values, where a darker color signifies increased correlation. For clarity, the Pearson correlation values have been transformed into an exponential scale in an attempt to place more emphasis on the differences among higher values. The options in this table were deliberately arranged in such a way that measures that are more correlated with each other appear consecutively. A high value between measure x and y corresponds to a high likelihood that similar POIs in respect of measure x will also be similar with respect to measure y.

	mi	aa	is	ov	ku	ka	cos	ρ	f1	jac	sr8	sr6
mi	1.00	0.67	0.83	0.61	0.79	0.79	0.84	0.84	0.80	0.80	0.57	0.55
aa	0.67	1.00	0.98	0.81	0.83	0.30	0.80	0.80	0.75	0.72	0.61	0.58
is	0.83	0.98	1.00	0.84	0.86	0.32	0.82	0.82	0.77	0.72	0.64	0.61
ov	0.61	0.81	0.84	1.00	0.97	0.64	0.83	0.83	0.70	0.63	0.58	0.54
ku	0.79	0.83	0.86	0.97	1.00	0.89	0.93	0.93	0.84	0.79	0.73	0.69
ka	0.79	0.30	0.32	0.64	0.89	1.00	0.97	0.98	0.92	0.92	0.87	0.87
cos	0.84	0.80	0.82	0.83	0.93	0.97	1.00	1.00	0.98	0.94	0.86	0.83
ρ	0.84	0.80	0.82	0.83	0.93	0.98	1.00	1.00	0.98	0.94	0.86	0.83
f1	0.80	0.75	0.77	0.70	0.84	0.92	0.98	0.98	1.00	0.98	0.90	0.87
jac	0.80	0.72	0.72	0.63	0.79	0.92	0.94	0.94	0.98	1.00	0.96	0.95
sr8	0.57	0.61	0.64	0.58	0.73	0.87	0.86	0.86	0.90	0.96	1.00	0.99
sr6	0.55	0.58	0.61	0.54	0.69	0.87	0.83	0.83	0.87	0.95	0.99	1.00

Figure 4.14: Pearson correlation among the projections. The magnitude of the correlation is imprinted via the color shade of each cell; a darker color indicates higher correlation.

The results indicate that the measures are generally positively correlated, an observation that is expected as these quantities correspond to similarities. In a smaller scale, however, some pairs appear to experience very similar behavior while some are less correlated than others. In particular, 5 groups of related measures can be distinguished: $\{aa, is\}$, $\{ov, ku\}$, $\{ka, cos, \rho, f1\}$, $\{jac, sr6, sr8\}$ and mi by itself. These results are driven by the specific dataset that we use and, hence, cannot be generalized, but they provide insights in terms of the different properties of the similarity measures.

4.8 Conclusions

The purpose of this chapter was to study the potential of user generated Point of Interest lists in recommendation systems. We proposed a methodology that only takes into consideration the bipartite graph structure of the POIs-lists graph, that is present in the Foursquare lists feature, to develop a POI recommendation system based on user profiles. This was made possible via similarity criteria that operate under this bipartite graph and can formulate the similarity between pairs of POIs. Our assumption was that information about the POIs is encoded in the lists and, in particular, that similar POIs will be submitted under the same list. Our evaluation confirmed, up to a certain extent that this assumption is reasonable. Specifically, we performed a user survey using local volunteers and used the responses to drive our algorithms and the evaluation process. The results indicated a significant correlation between the output of our method and the responses of the locals, particularly during the evaluation of the system as a personalized POI recommendation system. The offline evaluation strongly confirmed our approach and was in agreement with the findings of the online experiments.

We conclude this chapter by arguing that social network analysis can be an effective way of obtaining information about how users perceive POIs. Our methods took advantage of large amounts of publicly available user generated content to extract useful conclusions about the relationships among POIs, are very easy to reproduce and can be performed in real time. Through this work we hope to inspire the use of lists in further research as we believe the amount of useful information in this structure is profound and its potential is not limited to a recommendation or a points-of-interest perspective.

Future research on POI lists should focus on the extension of the graph structure as a way to leverage more available user generated data. A possible expansion is the incorporation of users in the graph, where they can be related to other entities as list creators or contributors. This scheme can be modeled as a tripartite graph of users, lists and POIs and can be processed using specific graph theoretic notions. The incorporation of users can expand the approach into possibly a user-centric method, for example based on user similarities or even in combination with a user relationship method such as collaborative filtering. Furthermore, while in this work we made use of the lists in the Foursquare LBSN, we believe our methods can be applied to any portal that

utilizes user generated lists, which might not necessarily be POI-related. For example, Amazon users can specify list of products (Listmania), and IMDb users can define their own lists of entities like movies and actors. In addition, Twitter has a lists feature via which users can group other profiles under the same combined feed. Similarly to our lists, these platforms share characteristics with respect to the physical interpretation of lists and, often, it is reasonable to consider them as groups of related entities. It is a question of future research to investigate the applicability of our approach to these or other platforms.

Chapter 5

Whole Sampling Generation of Scale-Free Graphs

This chapter presents the development of a new class of algorithms that accurately implement the preferential attachment mechanism of the Barabási–Albert (BA) model to generate scale-free graphs. Contrary to existing approximate preferential attachment schemes, our methods are exact in terms of the proportionality of the vertex selection probabilities to their degree and run in linear time with respect to the order of the generated graph. Our algorithms are based on a principle of random sampling which is called *whole sampling* and is a new perspective for the study of preferential attachment. We show that they obey the definition of the original BA model that generates scale-free graphs and discuss their higher-order properties. Finally, we extend our analytical presentation with computer experiments that focus on the degree distribution and several measures surrounding the local clustering coefficient. The results presented in this chapter are available in Stamatelatos and Efraimidis (2021b).

5.1 Introduction

The Barabási–Albert (BA) model (Barabási & Albert, 1999) is a growing preferential attachment mechanism that dictates the rules of connections among vertices when newborn nodes enter the network. Specifically, it requires that we select m vertices from the graph population when a newborn node enters the network in such a way that the probability of selecting a vertex is proportional to its degree. By repeating this process, the model results in the generation of a scale-free graph of order n.

Despite its widespread use the model was shown to have ambiguities in Stamatelatos and Efraimidis (2021a) regarding the exact node selection process during the addition of a newborn node. In that work, the tight relation between the BA model and the problem of weighted random sampling is being established, considering that selecting m vertices

from a population of n vertices where the probability of selections are proportional to the vertex degrees is a WRS problem. Considering this relation, the distinction on first and higher order inclusion probabilities is made and the impact of each type is quantified in experimental or analytical settings.

Initially, regarding the first order inclusion probabilities, it is stated that the original definition of the BA model is not perfectly clear on whether the probabilities that are proportional to the degree of the vertices are the inclusion, the selection, the independent or even something entirely different (also see Section 2.6.3). This is indirectly specified in the master equation approach proof of the power law degree distribution behavior of the method in Albert and Barabási (2002), where it is assumed that these refer to the inclusion probabilities, making the WRS scheme employed having the strπps property. The impact of schemes with different first order inclusion probabilities is studied experimentally to show quantifiable differences in the resulting degree distribution of the graph, the probabilities of individual nodes to occupy specific ranks in the degree ranking and the overthrow probabilities, which refer to the probability of individual nodes to occupy the top rank in the degree ranking.

In the same work, a more subtle ambiguity is being identified regarding the higher order inclusion probabilities of groups of vertices to be included together in the selection of m vertices when a newborn node enters the network. The definition of the model does not specify these probabilities and, while the degree distribution is not impacted, other, second order properties of the graph may be impacted. The impact is shown analytically using a counterexample and experimentally using visualizations of the projections of the resulting graphs. The projections highlight the second order properties of the graph (i.e., common neighbors instead of presence/absence of edge) and are used in the analysis of real social networks with significant success (Chapters 3 and 4).

The analysis in Stamatelatos and Efraimidis (2021a) shows that the BA model is not a single, well defined, model but a family of scale-free graph generation models that result from a growing scheme that satisfies the strπps property. Therefore, we define a round of the BA model as a strπps selection scheme without replacement and with constant sample size:

Definition 1. *(A round of the BA model) A new node is inserted into the graph. Then m distinct existing nodes are randomly selected such that the first order inclusion probability of every vertex is proportional to its degree. Finally, the degrees of the vertices are updated after the selection of m vertices is complete.*

The definition implies that the update of the degrees cannot occur during the selection process. The reader may refer to Stamatelatos and Efraimidis (2021a) for more information about the application of first and higher order inclusion probabilities in the BA model and how they impact the generators. It is worth noting that the arguments made in that work regarding the relation between the WRS problem and the preferential attachment mechanism are also relevant to other fields too, for example

the node2vec vertex embedding method (Grover & Leskovec, 2016) utilizes a form of negative sampling where vertices are selected with probability proportional to their degree (Armandpour et al., 2019), pointing to an application of WRS.

Current state of the art preferential attachment models are typically efficient in terms of their running time but they are not equivalent to the exact model described by Barabási and Albert; they do not refer to the same simple graph without multiple edges, or the probability model employed is only an approximation of the original scheme.

One of the most popular theoretical models in the literature is the model of Bollobás (Bollobás et al., 2001), which employs a probability model that guarantees strict proportionality but results in a multigraph. Due to its simplicity and the rigorous analysis made in this work about various properties of this network, the model has since been adopted in the literature. Another example of a multigraph generator is mentioned in Van Der Hofstad (2016, Chapter 8), where the edges are added with intermediate weight updates with replacement, a scheme that results in possible multiple edges per node pair.

Other studies correctly treat the BA model as a simple graph but with a probability selection scheme that is only an approximation of the original model. For example, Hadian et al. (2016) define a simple graph generator but the probabilities of node inclusions are not exactly proportional to their degree due to rejections. This difference has been explained further in Stamatelatos and Efraimidis (2021a), where the distinction of inclusion and selection probabilities is made. N. Berger et al. (2014) attempt to make a distinction about different probability schemes (independent, conditional, sequential) but also refer to the Bollobás multigraph model.

Despite the models mentioned previously being both efficient and rigorously studied, they do not strictly abide by the definition of the original BA model. The definition requires a sampling scheme without replacement that generates simple graphs, and the inclusion probability of a vertex is strictly proportional to its degree. These requirements can be summarized into the *strπps* random sampling scheme, which refers to a weighted random sampling design without replacement with inclusion probabilities strictly proportional to degree, and is also known as the *ratio estimator property* (Brewer & Hanif, 1983, Section 1.4).

In this chapter, we present a new class of algorithms that obey the definition of the BA model strictly, both with respect to the type of the output graph and the interpretation of the probabilities being employed in the preferential attachment step being exactly proportional to the node degrees. Our algorithms also run in linear time with respect to n, or in constant time for each time step, i.e., for each new node. It is worth noting that it is trivial to apply any *strπps* sampling method on each time step independently but that would result in quadratic complexity for the whole process. The computational complexity of the algorithm is of great importance since the sizes of the generated scale-free graphs are often very large, up to hundreds of thousands or even millions of nodes.

To our knowledge, up to now there was no efficient sampling algorithm for running each step in $\mathcal{O}(1)$ with respect to n. This could be a reason why current software libraries, e.g. NetworkX (Hagberg et al., 2008) or iGraph (Csardi, Nepusz, et al., 2006), have opted for a related efficient weighted random scheme that only approximates the BA definition. To the best of our knowledge this is the first time that both a strict interpretation of the inclusion probabilities and a linear running time are satisfied.

The basic principle of our algorithms can be demonstrated in the simple case where $m = 2$. During each time step, a random edge is selected and the newborn node is connected with the ends of that edge creating a triangle. The process is repeated until the graph is of the desired size. It is easy to show that on this sampling scheme, where each edge is guaranteed to connect two different nodes, each vertex exists in the edge set as many times as its degree and, therefore, its inclusion probability on any time step is exactly proportional to its degree, satisfying the *strπps* scheme. This concept of random sampling where the sample space is computed and maintained ahead of time so that a random sample can be generated in constant time is known as *whole sampling* (Brewer & Hanif, 1983, Section 1.7.1). Whole sampling is a new perspective of the algorithmic study of preferential attachment, which until now is based on consecutive node selections to execute a time step. The idea of sampling an edge has been mentioned before in Batagelj and Brandes (2005), where the authors claim that

> in a list of all edges created thus far, the number of occurrences of a vertex is equal to its degree, so that it can be used as a pool to sample from the degree-skewed distribution in constant time.

Based on the principle of this simple model, we generalize its operation for $m > 2$ by taking advantage of the operation of Jessen's whole sampling algorithm (Jessen, 1969), a weighted *strπps* random sampling scheme, and adjust its operation to fit the preferential attachment problem by introducing auxiliary data structures. Similarly to the simple $m = 2$ model where selections of pairs of nodes are made, our generalization allows the selection of m-tuples by parallelizing the list of edges with a list of hyperedges of an implicit m-uniform hypergraph. We prove that our approaches satisfy the *strπps* model and show their running time to be linear with respect to n.

5.2 Algorithms

In this section, we present the whole sampling algorithms for generation of scale-free graphs. We begin the presentation of our methods with the simple $m = 2$ algorithm that was described in Section 5.1; we label this algorithm SE-A (Section 5.2.1). Although this algorithm works only for $m = 2$, it demonstrates the basic principle of our approach. A generalization is then given for $m > 2$ as the abstract algorithm SE (Section 5.2.2) that utilizes an auxiliary data structure H that resembles a m-uniform hypergraph. Due to the generalization being abstract, we propose two possible implementations

that achieve the generation of scale-free graphs in slightly different ways. Algorithm SE-B (Section 5.2.3) is a more minimalistic approach towards the growth of the H data structure while algorithm SE-C (Section 5.2.4) performs more work, while still maintaining linear complexity, in an attempt to reduce correlations among tuples of vertices in the resulting graph. Both SE-B and SE-C reduce to algorithm SE-A when $m = 2$. We prove that both algorithms are correct with respect to the *strπps* probability model and have linear time worst case complexities.

Each algorithm is described as a growing process starting from an initial graph $\mathcal{G}_0$ until it reaches the desired order n. For simplicity, it is assumed that the starting graph is connected, otherwise unconnected graphs may be generated. A discussion surrounding the properties of initial graphs is given in Section 5.2.5.

5.2.1 Algorithm SE-A

The operation of algorithm SE-A can be summarized via its growth function. During the addition of one newborn node, one uniformly random edge is selected from the existing edge set and the new vertex is connected to the endpoints of that edge. A high level sketch of SE-A appears in Algorithm 2. All following algorithms are assumed to have an implicit random number generator input.

Algorithm 2 Algorithm SE-A (high level sketch)

Input An initial connected graph $\mathcal{G}_0(V_0, E_0)$ containing at least one edge and the desired number of vertices n

Output A scale-free graph $\mathcal{G}(V, E)$

1: $(V, E) \leftarrow (V_0, E_0)$
2: **while** $|V| < n$ **do**
3: select one uniformly random edge $e = (u, w) \in E$
4: add new vertex v to V
5: add new edges (v, u) and (v, w) to E
6: **return** $\mathcal{G}(V, E)$

The intuition of the method is that a vertex exists in the edge set as many times as its degree and no edge can contain the same vertex more than once. As a result, during a single uniform random edge selection, the inclusion probability of a vertex with degree d at any time is $d/|E|$. Therefore, the probability of a vertex with degree d gaining an edge at any time after a newborn node has been added is exactly proportional to its degree, which proves the following theorem regarding the correctness of the algorithm:

Theorem 1. *Algorithm SE-A satisfies the strπps probability scheme and generates a simple graph according to the definition of the BA model.*

Complexity

The work performed by algorithm SE-A during its growth function is one random edge selection, one vertex addition and 2 edge additions. Therefore, the whole process requires $\sim n$ vertex additions, $\sim 2n$ edge additions and $\sim n$ random variates, leading to the following theorem:

Theorem 2. *Algorithm SE-A runs in time* $\Theta(n)$ *to create a graph of n vertices.*

Here the tilde symbol is set to mean asymptotic equivalence (Bruijn, 1981, Section 1.4). Algorithm SE-A does not require any additional memory other than the output graph and doesn't use auxiliary data structures.

Clustering Coefficient

The simplicity of algorithm SE-A also allows us to analytically study several properties surrounding the local clustering coefficient. For the following analysis, it is assumed that the initial graph $\mathcal{G}_0$ consists of two vertices and 1 edge, which is the smallest graph from which the operation can start. First, we state the following theorem regarding the correlation of the local clustering coefficient with the degree:

Theorem 3. *The local clustering coefficient of a vertex with degree* d *is* $2/d$ *at any time of the generation process.*

Proof. The local clustering coefficient of a vertex i at time t is

$$C(d) = \frac{2E_i}{d_i(d_i - 1)}, \tag{5.1}$$

where E_i is the amount of edges among i's neighbors and d_i the degree of i. Therefore, when i enters the network, its local clustering coefficient is 1, because $d = 2$ and $E_i = 1$. If i does not acquire any new edges, its local clustering coefficient will not change, because none of the quantities E_i or d_i will change, since no edges are created among existing vertices. If i obtains one edge, both its degree and E_i will be increased by 1, because new vertices only connect to existing edge's endpoints. Thus, d and E_i are always connected via the formula $d = E_i + 1$. Replacing in Equation 5.1 gives

$$C(d) = \frac{2(d-1)}{d(d-1)} = \frac{2}{d}. \tag{5.2}$$

It is easy to see that the formula also holds for the 2 vertices in the initial graph after the third node has entered. □

Theorem 3 shows that the local clustering coefficient of a vertex depends only on its degree. As a result, the limit of the average local clustering coefficient can now be derived based on the expected power law degree distribution:

Theorem 4. *The limiting average local clustering coefficient of a graph produced using the SE-A algorithm is* $2\pi^2 - 19$.

Proof. The degree distribution of a graph generated using the *strπps* method when $n \to \infty$ is given in Albert and Barabási (2002, Equation 90):

$$P(d) = \frac{2m(m+1)}{d(d+1)(d+2)}, d \geq m. \tag{5.3}$$

By combining (5.2) and (5.3) and setting $m = 2$, we can get the average local clustering coefficient:

$$C_{avg} = \sum_{d=2}^{\infty} (2/d)P(d) = \sum_{d=2}^{\infty} \frac{24}{d^2(d+1)(d+2)} = 2\pi^2 - 19 \approx 0.73921. \tag{5.4}$$

□

Additionally, the local clustering coefficient distribution can also be formulated. Exchanging d with $2/c$ and setting $m = 2$ in Equation 5.3, we get the local clustering distribution P_c:

$$P_c(c) = P(2/c) = \frac{6c}{(2/c+1)(2/c+2)}. \tag{5.5}$$

The value of the average limiting local clustering coefficient produced by algorithm SE-A is inline with empirical observations of real social networks (Albert & Barabási, 2002, Table I). Thus, the SE-A mechanism can simulate features of real social networks beyond the power law degree distribution.

Following the analysis above, although the degree distribution does not have a finite variance, we can deduce the variance of the clustering coefficient:

$$\begin{aligned}
\sigma^2[C] &= E[C^2] - (E[C])^2 \\
&= \sum_{d=2}^{\infty} (2/d)^2 P(d) - C_{avg}^2 \\
&= 24\zeta(3) - 4\pi^4 + 70\pi^2 - 330 \\
&\approx 0.08531,
\end{aligned}$$

where $\zeta(3)$ is the constant

$$\zeta(3) = \sum_{k=1}^{\infty} \frac{1}{k^3}.$$

The operation of the SE-A algorithm can so be parallelized with Holme and Kim (2002), where the clustering coefficient is tunable by setting a probability of creating a triangle when new vertices enter the network. Here, this probability is 1 because the new vertex always connects with two endpoints of the same edge. As a result, the number of triangles in the final graph $\mathcal{G}$ is $tri(\mathcal{G}) = tri(\mathcal{G}_0) + n - 2 \sim n$.

Finally, it is worth mentioning that the case of $m = 1$ is not discussed here as there is no distinction between the interpretation of the proportionality of the probabilities and, hence, existing mechanisms obey the *strπps* model for this particular case.

5.2.2 Generalized Abstract Algorithm SE

Algorithm SE-A can be extended for $m > 2$ using a whole sampling method that perfectly fits the constrains and requirements of this problem and is due to Jessen (1969). Jessen's method builds a sample space (called *tableau*) according to the given inclusion probabilities as m-tuples of elements in an iterative way. Each element is assigned a balance quantity proportional to its inclusion probability which is reduced each time the element is used in a sample; the method terminates when all balances are depleted. It is then possible to select one *strπps* sample of m elements in constant time.

Here, we exploit the constant time selection and the growing nature of Jessen's method to define the abstract algorithm SE. Algorithm SE maintains a tableau of possible samples as an auxiliary data structure H, which comprises a list of m-tuples such that each node exists in as many tuples as its degree. Updating this data structure when newborn nodes enter the network can be performed by increasing the balance of the newborn node and the selected vertices in the tableau without having to repeat the process. We note that for algorithm SE-A, the H data structure is equivalent to the edge set of the network and, hence, not required concretely. The H data structure resembles a m-uniform hypergraph, where each of the m-tuples of the list represents one hyperedge. The nature of the process allows multiple copies of the same hyperedge in H, similarly to Jessen's method allowing for the same row in the tableau. As a result, H may represent a non-simple hypergraph where repeated edges are possible. Note that even though the hypergraph is non-simple, the m-uniform property assures that each hyperedge contains exactly m distinct vertices. Therefore, no loops are permitted. In the rest of the document, we refer to H and its contents in the hypergraph terminology. A high level sketch of this abstract algorithm appears in Algorithm 3.

Algorithm SE, as defined here, is abstract with respect to the `update` function, which can be implemented in numerous ways. This function corresponds to the maintenance of the H hyperedge set when newborn nodes enter the network and is required to satisfy two invariants after the `update` function returns:

1. Each vertex may only exist at most once in a hyperedge.
2. Each vertex participates in as many hyperedges as its degree in $\mathcal{G}$.

These invariants guarantee the correctness of any algorithm, as they are the only necessary conditions for the sampling scheme to be *strπps*. These requirements are implicitly satisfied in the SE-A algorithm since the H data structure is identical to the edge set. The invariants can be simplified by stating that

Algorithm 3 Abstract Algorithm SE (high level sketch)

Input An initial connected graph $\mathcal{G}_0(V_0, E_0)$, the desired number of vertices n and the desired number of edges added per step m

Output A scale-free graph $\mathcal{G}(V, E)$

1: $(V, E) \leftarrow (V_0, E_0)$
2: init(H, $\mathcal{G}_0$) ▷ initialize H based on $\mathcal{G}_0$
3: **while** $|V| < n$ **do**
4: select one uniformly random hyperedge $e = (e_1, e_2, \dots, e_m) \in H$
5: add new vertex v to V
6: add new edges $(v, e_1), (v, e_2), \dots, (v, e_m)$ to E
7: update(H, v, e) ▷ update H based on v and the contents of e
8: **return** $\mathcal{G}(V, E)$

> H must be a (possibly non-simple) m-uniform hypergraph with the same degree sequence as $\mathcal{G}$.

A general operation of the `update` function is to handle the updating of the H data structure based on the newborn node addition. In particular, the vertices $e_1, e_2, \dots, e_m$ and m copies of the newborn node v have to be added in H. These $2m$ items imply the addition of 2 new hyperedges in H. Since no more than 2 copies of v can be added in 2 hyperedges, previously added hyperedges need to be adjusted as well to satisfy the invariants. Two possible methods of achieving this are described in Sections 5.2.3 and 5.2.4.

The `init` function represents the initialization of the H data structure so that the invariants are satisfied during the start of the process. Directly following the definition of the invariants, it can be seen that the initial graph $\mathcal{G}_0(V_0, E_0)$ needs to satisfy the divisibility $2|E_0|/m$ and no vertex can have degree higher than $2|E_0|/m$ for the H data structure to be feasible. The requirements of the initial graph $\mathcal{G}_0$, which are omitted from Algorithm 3 for brevity, are discussed in more detail in Section 5.2.5.

The `init` function is marked as abstract because it is possible to be implemented in various different ways. Here, we describe one possible implementation that distributes the vertices in V_0 randomly throughout H. The method, which we call *random systematic partitioning*, has been influenced by the random systematic sampling method (Goodman & Kish, 1950), which we here adjust to partition the items instead of sampling them. This algorithm, which –to our knowledge– does not seem to have been described before in the literature, might be of independent interest. Random systematic partitioning accepts a bag of elements $x_1, x_2, \dots, x_n$ where each element x_i is characterized by its frequency d_i (the degree in this context). The goal of the algorithm is to partition the bag into s sets of m elements, where no vertex can exist more than once in any of those s sets. For a feasible outcome, it must hold that $s \cdot m = \sum_{i=1}^{n} d_i$ and the maximum frequency cannot be higher than s. If the frequencies are not known in advance they

can be created in one pass over the population.

A high level sketch of random systematic partitioning can be seen in Algorithm 4. The algorithm initially shuffles the unique x values and expands them into their frequencies into an implicit $s \times m$ matrix written by row. The output of the operation is the transpose of this matrix; each of the s rows represents one group of the partition. The computational complexity of random systematic partitioning is $\Theta(sm)$ because of the encapsulated loops and is formalized in the following theorem:

Theorem 5. *Given a multiset with $s \cdot m$ items with at most s repetitions of each item, random systematic partitioning runs in time $\Theta(sm)$, and partitions the multiset into s sets of m items each.*

Algorithm 4 Random Systematic Partitioning Algorithm (high level sketch)

Input A bag of elements $x_1, x_2, \ldots, x_n$ where each element i appears with frequency d_i, the desired number of sets s and the desired number of elements in a set m

Output A partitioning H of the input bag into s sets of m elements randomly distributed across the data structure

1: $H \leftarrow$ array of s empty sets
2: $k \leftarrow 1$
3: shuffle x
4: **for all** $i \in [1, n]$ **do**
5: **for** d_i times **do**
6: add i to H_k
7: $k \leftarrow (k \mod s) + 1$
8: **return** H

5.2.3 Algorithm SE-B

Algorithm SE-B is an implementation of algorithm SE with a minimal approach into implementing the `update` function. The main issue of distributing the m copies of the newborn node v is addressed by inserting one copy into each of the 2 new hyperedges and $m - 2$ copies into previously inserted hyperedges. In the latter case, one node from each of these $m - 2$ is swapped back into the new hyperedges. A high level sketch of the `update` function of algorithm SE-B is shown in Algorithm 5.

First, the two new hyperedges h_x and h_y are initialized with the vertices from the randomly selected hyperedge e. Although this choice does not impact the correctness of the algorithm and can be executed arbitrarily, a sensible option is a half split, or a near-half split if $|e|$ is odd. One copy of the newborn node v is then added to each of the h_x and h_y hyperedges as it cannot have been previously contained in either. The algorithm then selects $m - 2$ existing hyperedges to perform the swap of the $m - 2$ remaining copies of v. The selection of these hyperedges is also irrelevant to the correctness of

Algorithm 5 Algorithm SE-B – update function (high level sketch)

Input The existing hyperedge list H, the newborn node v and the selected hyperedge e

Output The new state of the hyperedge list H

initialize two new hyperedges h_x and h_y
divide the vertices of e in h_x and h_y
add v in h_x and v in h_y
select $m-2$ hyperedges from H: $h_1, h_2, \ldots, h_{m-2}$
for all $i \in [1, m-2]$ **do**
    $h_c \leftarrow$ a non-empty hyperedge in $\{h_x, h_y\}$
    find an element w of h_i not present in h_c
    add w in h_c and replace it with v in h_i
add h_x and h_y on H
return H

the algorithm and can be performed arbitrarily but it is generally desirable or useful to randomize the selection. One possible algorithm is shown in Batagelj and Brandes (2005, Section II.B, Alg. 3) that operates using a virtual shuffle and performs $m-2$ selections without the possibility of collisions. This algorithm has the property that all higher-order inclusion probabilities are equal, i.e. all possible $m-2$ groups of hyperedges are equiprobable to be selected. It is worth noting that e might be one of these $m-2$ hyperedges. The algorithm selects one node from each of these $m-2$ hyperedges that is not already present in either h_x or h_y and swaps its value with v. This selection can also be done in different ways but one option consistent with the choices made previously is to traverse an existing hyperedge in random order until a node is found not to be contained in the new hyperedge. Finally, the new hyperedges h_x and h_y are appended into H. Note that Algorithm SE-B reduces to Algorithm SE-A for $m = 2$, as no exchanges are taking place. A diagram of the operation of Algorithm SE-B for $m = 4$ is shown in Figure 5.1.

Algorithm SE-B satisfies the correctness invariants, as during the updating of the H hypergraph, the newborn node that has degree m gains exactly m hyperedges and each node in e, whose degree is increased by 1, gains one more hyperedge. No additional insertions or removals are performed, except swaps, leading to the following theorem:

Theorem 6. *Algorithm SE-B satisfies the strπps probability scheme and generates a simple graph according to the definition of the BA model.*

The complexity of algorithm SE-B can be derived from the complexity of the `update` function in Algorithm 5. The initial steps are operations that can be performed in time proportional to m, including the selection of the $m-2$ existing hyperedges. In order to fill h_x and h_y, there are $2 \cdot (m/2 + (m/2+1) + (m/2+2) + \ldots + (m-1)) = \Theta(m^2)$

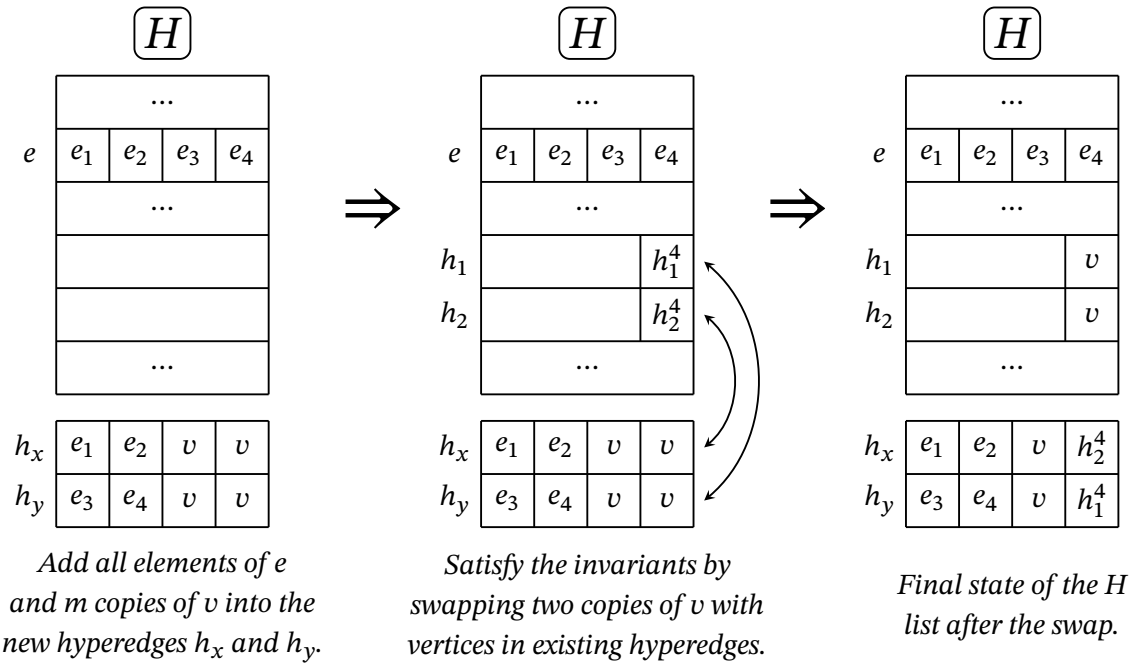


Figure 5.1: Demonstration of the operation of Algorithm SE-B for $m = 4$. On the left, the contents of e and m copies of v are added in the new hyperedges temporarily. This step is not explicit in the algorithm, as normally only 2 copies of v would be added, and it simply aids the visualization. In the middle, $m - 2$ existing hyperedges have been selected and one value of each (h_1^4 and h_2^4) is identified as swappable with the new hyperedges. The state to the right is the final state of the H data structure after swapping these values. The new node v is in m hyperedges, whereas nodes $e_1 \ldots e_4$ in one more hyperedge than before.

operations required in the worst case when every element checked in the existing hyperedges except for the last exists in either h_x or h_y. This leads to the following theorem:

Theorem 7. *Algorithm SE-B runs in time $\mathcal{O}(nm^2)$ to create a graph of n vertices.*

Despite algorithm SE-B running in linear time with respect to n, the nm^2 complexity of the worst case is unlikely to occur in a typical instance of the problem. Given that as n increases, the probability of collisions when finding elements in the existing hyperedges not already present in h_x and h_y is being reduced, we conjecture that the average case complexity of algorithm SE-B is $\mathcal{O}(nm)$. This hypothesis is supported by experimental observations but should be pursued in future work.

5.2.4 Algorithm SE-C

Algorithm SE-B works correctly with respect to the probability model involved but has one inherent property regarding the higher-order inclusion probabilities in the preferential attachment step, which refers to the probability of certain groups of vertices to acquire common neighbors during the growing process. Specifically, groups of vertices that have been selected together in the past are more likely to also be selected together in the future, since the nodes inside e are used to populate the new hyperedges. This behavior is sometimes desired, for example real social networks are not typically uncorrelated and have some degree of underlying structure. In this section, we describe algorithm SE-C, which can be used in situations where the above behavior is not desired. Unlike algorithm SE-B, algorithm SE-C is tunable with respect to the randomization and shuffling of the vertices inside the H data structure and can potentially minimize the effects of the higher-order inclusion probability bias.

The core idea of algorithm SE-C is, instead of swapping only a single element of each of the $m-2$ existing hyperedges to replace v, to shuffle all copies inside the $m-2$ existing hyperedges as well as h_x and h_y (m^2 elements in total). Random systematic partitioning (Algorithm 4) fits this concept perfectly, as it is guaranteed that no vertex will have more copies than the number of hyperedges (v being the maximum with m copies).

A high level sketch of the `update` function of algorithm SE-C is shown in Algorithm 6. Following the scheme of random systematic partitioning, the contents of the $m-2$ existing hyperedges and e as well as m copies of v are inserted into the partitioning algorithm and the resulting groups are replacing their old records, comprising the new value of the H hypergraph. It is worth noting that the shuffling performed by algorithm SE-C can be tuned by increasing the number of hyperedges inserted into random systematic partitioning, for example $2m$ (2 new and $2m-2$ existing) instead of m.

Algorithm SE-C performs the same amount of additions as Algorithm SE-B. Therefore,

Algorithm 6 Algorithm SE-C – update function (high level sketch)

Input The existing hyperedge list H, the newborn node v and the selected hyperedge e

Output The new state of the hyperedge list H

1: initialize empty array A
2: select and remove $m-2$ hyperedges from H: $h_1, h_2, \ldots, h_{m-2}$
3: add $m \cdot (m-2)$ elements from $h_1, h_2, \ldots, h_{m-2}$ into A
4: add m copies of v into A
5: add all elements of e into A
6: perform random systematic partitioning on A with $s = m$
7: add the m sets of A to H
8: **return** H

there is no change in the correctness in relation to algorithm SE-B:

Theorem 8. *Algorithm SE-C satisfies the strπps probability scheme and generates a simple graph according to the definition of the BA model.*

Regarding its complexity, Algorithm SE-C is bounded by the random systematic partitioning that is applied on m^2 elements and, hence, considering Theorem 5, its running time is proportional to nm^2. Unlike Algorithm SE-B, the running time is not impacted by the random number generator and its asymptotic performance is always proportional to nm^2:

Theorem 9. *Algorithm SE-C runs in time $\Theta(nm^2)$ to create a graph of n vertices.*

As a closing remark, it is interesting to note that algorithm SE-C demonstrates the close association between preferential attachment and the random sampling problem. In fact, three different random sampling methods are involved in the design of algorithm SE-C to solve a preferential attachment problem:

1. One sampling algorithm to select $m-2$ existing hyperedges from the population of hyperedges in H.

2. Random systematic partitioning, which is influenced by systematic random sampling and is used to both initialize the H array from $\mathcal{G}_0$ and to shuffle the node copies inside the m hyperedges.

3. Jessen's whole sampling method to update the H array in such a way that the inclusion probabilities are always proportional to the degrees of the vertices.

In this chapter, we exploit this relation in order to develop an implementation of the growing preferential attachment mechanism that perfectly fits the requirements of the *strπps* probability scheme. Future integration between these two problems should also be pursued in the future.

5.2.5 Discussion: The Initial Graph

In this section, we discuss the options for the initial graph $\mathcal{G}_0(V_0, E_0)$ that can be used in our algorithms, explain its requirements that were previously omitted for brevity, and propose methods to satisfy them.

The global features of the scale free graph are expected to be independent of the initial graph, as the BA model is typically regarded a stationary distribution model. However, the initial graph state influences features of the first nodes, for example the probability that a specific node will become the heaviest node in the social network. For consistency and completeness, we note that the initial graph constitutes a state of our methods (the first state) and, as such, needs to satisfy the invariants of Section 5.2.2 in order for the transformation into the hypergraph H to be feasible. In particular:

1. The number of edges $|E_0|$ in the initial graph needs to satisfy the divisibility $2|E_0|/m$.
2. No vertex can have degree higher than $2|E_0|/m$.

It is worth noting that the complete graph of m vertices, which is a typical initial graph used in the BA model, satisfies both of these conditions without any processing required.

Regarding requirement (2), and assuming that requirement (1) is satisfied, a vertex may not have degree that is bigger than $2|E_0|/m$, otherwise the number of hyperedges in the H data structure will not be enough for the copies of this vertex; at least one hyperedge would have to contain multiple copies, which is not allowed. For example, in the star graph of 5 vertices and $m = 4$, the center node has degree 4 while the sum of the degrees is 8. Thus, there are 2 hyperedges in H but the center node needs to have 4 copies in H, which is impossible. This situation highlights the inherent issue of infeasibility in random sampling when the *str*π*ps* model is used Efraimidis, 2015. In the previous example, the inclusion probability of the center node is $(4/8) \cdot 4 = 2$ (200%). A straightforward approach is to accept the fact that the probabilities are infeasible and to bound all infeasible probabilities to be at most 1, until the probabilities gradually become feasible as the number of nodes n increases.

Regarding requirement (1), considering that H needs to contain an integer amount of hyperedges, it follows that $2|E_0|$ needs to be divisible by m. For example, a complete graph of 6 nodes for $m = 4$ does not satisfy this property (30 node copies are not divisible by 4). In the rest of this section, we discuss two methods to address the limitations imposed by requirement (1), namely forcing the number of edges to a specific value that does not oppose the requirement and introducing a multiplication factor that enlarges the count of all entries of the problem.

Forcing the number of edges Initially, $\mathcal{G}_0$ can be transformed into a graph that satisfies requirement (1) by selecting an appropriate number of edges and using the

$\mathcal{G}(n, M)$ generator to produce the initial graph. The minimum number of edges in the initial graph such that it satisfies the requirement is

$$|E_0|_{min} = \frac{\text{lcm}(m, 2)}{2},$$

while the largest number of edges depends on $|V_0|$ and is

$$|E_0|_{max} = \left\lfloor \frac{|V_0| \cdot (|V_0| - 1)}{\text{lcm}(m, 2)} \right\rfloor \cdot \text{lcm}(m, 2).$$

Thus, it follows that

$$|V_0| \cdot (|V_0| - 1) - \text{lcm}(m, 2) \leq |E_0|_{max} \leq |V_0| \cdot (|V_0| - 1),$$

which implies that $|E_0|_{max}$ is within a margin of m or $2m$ of the edges of the complete graph with the same number of vertices that is often used as input. Naturally, the $\mathcal{G}(n, M)$ generator is still subject to requirement (2) and rejections should be used to ensure that.

Introducing a multiplication factor Another way to address the limitation caused by requirement (1) is to setup a factor of multiplication for the entire process. The multiplication factor is

$$\lambda = \frac{\text{lcm}(2|E_0|, m)}{2|E_0|}$$

and denotes the factor with which all node copies are multiplied with. Hence, for the initial graph, instead of inserting $2|E_0|$ entries in H, which might not be divisible by m, we are inserting $2\lambda|E_0| = \text{lcm}(2|E_0|, m)$ entries, which is divisible by m. Similarly, for the duration of the process, instead of inserting 2 hyperedges, we insert 2λ hyperedges, where the copies of the vertices are multiplied also by λ. While our algorithms are easy to be generalized to support this process and completely solve the limitation if such behavior is desired, this method costs in memory and design complexity and for most cases the solution of generating an initial $\mathcal{G}(n, M)$ graph with the closest number of acceptable edges should be preferred.

5.3 Experimental Approach

In this section, we present some computer experiments that highlight the behavior of algorithms SE-A, SE-B and SE-C in practical situations for finite n. We focus on two properties that typically arise in social network analysis of scale-free graphs: the degree distribution and properties surrounding the local clustering coefficient. We specify that for all experiments the version of the SE-C algorithm refers to the algorithm that shuffles m hyperedges (instead of more) while the random systematic partitioning algorithm is used for the initialization of the auxiliary hypergraph H. Finally, the split

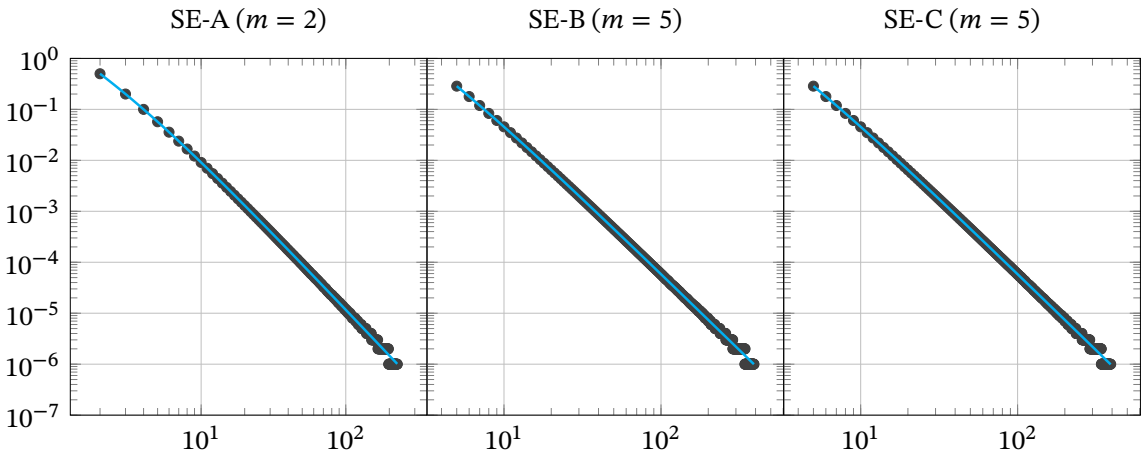


Figure 5.2: Degree distribution for the SE-A, SE-B and SE-C algorithms in a log-log plot for $n = 300\,000$. For the SE-A algorithm it is $m = 2$ while for SE-B and SE-C it is $m = 5$. The cyan lines that are rendered on top of the marks show the theoretical degree distribution. The close association between the theoretical expectation and the experimental models can be observed.

of e in the two new hyperedges is performed using the random method with equal split. The reference implementation of our algorithms is available online[1].

5.3.1 Degree distribution

For the experimental approach of the degree distribution of our algorithms we use $n = 300\,000$ in order to capture an approximation of the asymptotic state of the limiting distribution. By definition, the SE-A algorithm is only compatible with $m = 2$. For the SE-B and SE-C algorithms we use $m = 5$, as using $m = 2$ would result in identical behavior to SE-A, but within the limits of what is often used in practice. These are also the settings of m that are used throughout this section. In terms of the initial graph, the complete graph of $|V_0| = m$ is used.

Figure 5.2 shows the experimental degree distributions of the three algorithms; from left to right SE-A, SE-B and SE-C. The plots were generated using 10 000 iterations of the generation process to achieve statistical stability. The plots also contain the theoretical degree distribution (Equation 5.3) as a cyan line rendered on top of the data points. We note that, despite the theoretical distribution being a discrete probability distribution, it is rendered here as continuous in order to be visually distinguishable from the data points. The simulation shows that the resulting graphs are scale-free and an almost exact fit with the theoretical distribution. Although algorithms SE-A, SE-B and SE-C are different in their operation and their internal preferential attachment mechanism, they all result in scale-free distribution because they all satisfy the *strπps* property, as otherwise proven in Albert and Barabási (2002).

[1] https://github.com/gstamatelat/preferential-attachment-se

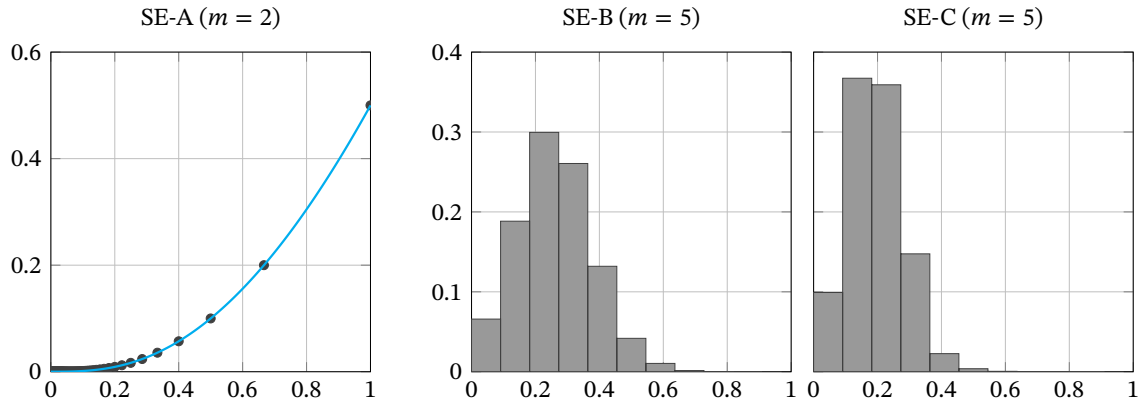


Figure 5.3: Local clustering coefficient distribution for the SE-A, SE-B and SE-C models. The settings refer to $n = 300\,000$. For the SE-A model, the theoretical local clustering distribution is rendered on top of the data points. The SE-B and SE-C methods are data binned in histograms in an attempt to reduce fluctuation noise levels. Besides the perceivable association of the SE-A distribution with its theoretical counterpart, an apparent difference among the distribution shapes of the methods is observed.

5.3.2 Local clustering coefficient distribution

The local clustering coefficient is typically used in social network analysis in order to study the degree to which nodes in a graph tend to cluster together. Initially, we experimentally show the local clustering coefficient distribution for our three models. While the distribution is analytically found for the SE-A model (Equation 5.5), the experiment provides insight about this distribution in the SE-B and SE-C models, which might be more difficult to gain analytically. It should be noted that, while the degree distribution of the three models is identical in terms of its shape according to the *strπps* property, the same is not guaranteed in the experiments regarding the local clustering coefficient that follow.

Figure 5.3 shows our experimental results. The same n and m settings were used as in the degree distribution experiment; we also performed 10 000 iterations here for statistical stability. For the case of SE-A, the theoretical local clustering coefficient distribution is displayed (lighter cyan line), with a strongly perceivable association with the experimental data points. In this case too, the theoretical distribution is displayed as a continuous distribution. For the SE-B and SE-C models, the distribution is more complicated and contains significant fluctuation noise; for this reason it is displayed here as a histogram of linearly binned data. The local clustering coefficient distributions of the SE-B and SE-C algorithms appear to be bitonic and resemble the log-normal distribution but this should be investigated further.

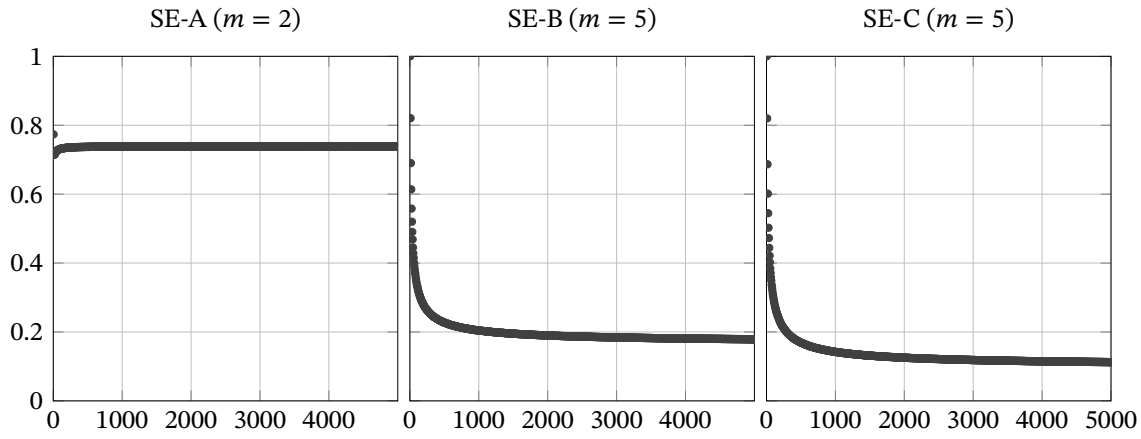


Figure 5.4: Average local clustering coefficient with respect to n for the SE-A, SE-B and SE-C models. The horizontal axis spans from 5 to 5 000. While SE-A immediately converges to its theoretical average, the SE-B and SE-C models have declining behavior.

5.3.3 Average local clustering coefficient

Another perspective of the clustering properties of our algorithms is the average local clustering coefficient with respect to its size n. While the results displayed in Figure 5.3 indicate that the average local clustering coefficient of SE-C is lower than this quantity for SE-B due to the distribution being more biased towards the lower x values, the plot does not show how the average is shaped over the duration of the generation process. For this analysis, we use the same settings for m as previously and increase n from 5 to 5000 in order to observe how the average local clustering coefficient behaves.

Figure 5.4 presents the results of the simulation of 1 000 iterations of the experiment. It can be observed that Algorithm SE-A converges fast to its theoretical expectation given in Equation 5.4. For the SE-B and SE-C there is an apparent declining behavior with respect to n which is due to the initial graph being a clique with average local clustering coefficient 1. The average local clustering coefficients of both algorithms appear to converge with Algorithm SE-C having smaller values for graph sizes n up to 5 000.

5.3.4 Degree correlation with local clustering coefficient

The convergence of the average local clustering coefficient of Figure 5.4 can be better understood via the correlation among the degree and the local clustering coefficient. The properties surrounding the connection between these two quantities for uncorrelated scale-free graphs have been studied before, for example in Bornholdt and Schuster (2003, Section 2.2), Colomer-de-Simon and Boguñá (2012) and Catanzaro et al. (2005). Due to our graph generators not being uncorrelated, we approach this association experimentally. The same settings of $n = 300\,000$ and $m = 2,5$ are used in this experiment as well.

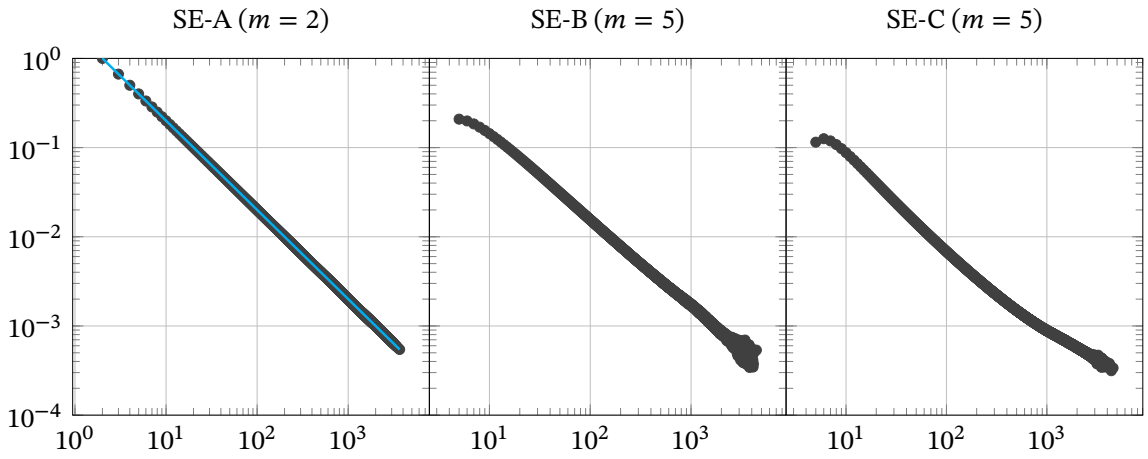


Figure 5.5: Correlation between the degree (horizontal axis) and local clustering coefficient (vertical axis) for the SE-A, SE-B and SE-C algorithms. The lighter cyan line that is rendered on top of the data points display the theoretical correlation for algorithm SE-A. An approximate power-law behavior can be observed for SE-B and SE-C.

Our results are illustrated in Figure 5.5. In this scatter plot, the horizontal component of each point represents the degree value and the vertical component represents the average local clustering coefficient of the vertices with that degree. The plots are averaged over the results of 20 000 iterations for the SE-B and SE-C figures and 1 iteration for the deterministic SE-A experiment. The theoretical correlation of the SE-A algorithm, given in Theorem 3 ($2/d$) is also presented in the figure as a continuous ligher cyan line. A declining behavior is observed for all three methods, which is consistent with previous findings of uncorrelated graphs (Bornholdt & Schuster, 2003; Catanzaro et al., 2005; Colomer-de-Simon & Boguñá, 2012). Moreover, the convergence of Figure 5.4 can be further explained. Following the process shown in Equation 5.4 and the declining nature of the degree-clustering correlation, it can be easily seen that the average local clustering coefficient converges to a non-zero constant for both SE-B and SE-C cases as $n \to \infty$.

5.3.5 Discussion: The higher-order case

The experimental simulations of this section also raise discussion regarding the high-order dynamics of the BA model and the preferential attachment mechanism in general. While the first-order properties define the probability that individual vertices are selected during a step of the preferential attachment process, the higher-order properties define the probability that groups of vertices are selected together during a single step of the growing process. The issue has been raised before (Stamatelatos & Efraimidis, 2021a), where it is claimed that the BA model is not one model but a family of models that respect the same first-order probability model but not necessarily the higher-order model. In the original works of Barabási and Albert, it is shown that the degree distribution,

and therefore the scale-free nature of the resulting graph, only depend on the first-order properties of the generator. Given that the scale-free nature of the BA model is regarded as the most interesting property of the model, the properties surrounding the higher-order case are less studied and known.

The algorithms SE-A, SE-B and SE-C presented in this chapter are examples of the distinction between the first and higher order probabilities of the preferential attachment mechanism. While their degree distribution is identical (assuming the same value for m), our experiments indicate a significant difference on their higher-order properties. One such property is related to the clustering coefficient and the probabilities of certain groups of vertices to gain a common neighbor on each of the preferential attachment steps. This phenomenon, in turn, impacts the probabilities of triangle formations or the degree at which certain groups of vertices form stronger clusters. This situation is more prevalent and extreme in the SE-A algorithm, where pairs of vertices that are not connected via an edge will never gain a common neighbor. The quantities surrounding the higher-order properties of different scale-free graph generators are worth investigating and positioned in relation to each other in the future.

5.4 Conclusions

In this chapter, we have utilized multiple random sampling schemes and methods to design a class of scale-free graph generator algorithms. Our models obey the dynamics of the preferential attachment scheme and the definition of the BA model. In particular, the algorithms are designed such that the inclusion probability of any vertex and at any time of the process is exactly proportional to its degree. This behavior is in contrast to existing methods where the inclusion probabilities are only approximately proportional to their degree. Our algorithms, that are primarily based on the concept of *whole sampling*, also run in linear time with respect to the desired graph size n. This is, to the best of our knowledge, the first time that both strict probability interpretation and linear complexity are achieved in the literature for the generation of scale-free graphs via the preferential attachment mechanism.

Our analysis started with algorithm SE-A, the simple case for $m = 2$ that demonstrates the principle of our approach, where one uniform random edge is selected and a newborn node is connected with its endpoints. The correctness of this algorithm was shown by the fact that a vertex exists in the edge set as many times as its degree. The generalization of SE-A, algorithm SE, is proposed, that uses an auxiliary data structure H that resembles a m-uniform hypergraph and works for $m \geq 2$. While the operation of this algorithm is abstract, the necessary invariants are defined that guarantee its correctness. Finally, algorithms SE-B and SE-C implement algorithm SE more concretely by either exchanging the necessary values in H to ensure the invariants are satisfied or completely shuffling parts of H respectively. Our computer simulation

experiments confirm the scale-free nature in the resulting graphs and raise interesting observations and future work directions about the higher-order properties surrounding the clustering features of the resulting scale-free graphs.

Chapter 6

Software Frameworks

In this chapter, two frameworks that have been developed in the context of this thesis are presented, namely `random-sampling` and `social-influence` libraries. In particular, `random-sampling` is a collection of implementations for reservoir sampling algorithms, both weighted and unweighted, where the memory consumed by each implementation is linear with respect to the size of the required sample. The `social-influence` library is a broader collection of tools and algorithms that range from the implementation of graph structures to tools for social behavior analysis or influence social models.

These frameworks were developed from scratch to close existing gaps in the open source scene regarding the topics discussed in this thesis. In particular, the concept of weighted weighted random sampling that is partially addressed with `random-sampling` appears to be creating confusion with respect to the interpretation of the weights. The development of these libraries allowed the experiments referenced in this thesis to be conducted, in most of the cases as the main component where they were tailored to the experimentation process and other times in secondary independent tools and scripts. Despite their tight relation with the subjects discussed in this thesis, their scope is more general and can be utilized for a broader set of applications.

The frameworks are open source and are hosted on Github[1], the largest platform of online code sharing and version control. They are written in Java 8 and are published in online software artifact repositories for anyone to import and use. Both frameworks are developed under the premise that they are efficient, fast and adhere to modern programming standards and practices. Their public API definitions are reasonable to the extent provided by the Java programming language, for which integration has been provided in a sensible way depending on the specific situation, for example the implementation of the `hashCode` method, the `equals` method, built-in object serialization and others.

In the following sections, detailed information about these frameworks is given in

[1] https://github.com

regards to their public API, their operation and internals. Use cases and examples of specific algorithms are also presented. The `random-sampling` framework is described in Section 6.1 and the `social-influence` library is described in Section 6.2.

6.1 The `random-sampling` Framework

The problem of random sampling, both unweighted and weighted, is a subject with significant role in this thesis. In particular, and as being discussed in Chapter 5, the preferential attachment mechanism is identified as an application of weighted random sampling. The implications of this observation is that very popular models, including the BA model, are open to interpretation with respect to the semantics of the proportionality of the probabilities involved. As a result, and as previously discussed, the literature in this subject appears fragmented and the connection of WRS with the preferential attachment mechanism is less known in this scientific field.

The consequences of these subjects in the open source software community are apparent. While there exist implementations in nearly all programming languages for the BA model, none of these implementations follow the exact preferential attachment scheme of the inclusion probability interpretation required to create scale-free graphs whose degree distribution follow the Yule distribution as originally proposed in the work of Barabási and Albert. In addition, the distinction among the different probability models of the WRS designs is also less known in the open source software community where each design appears to give a different interpretation.

The `random-sampling` package was made to fill these gaps and to give attention to both of the subjects of the differences among the WRS designs as well as their inherent impact on the preferential attachment mechanism. Furthermore, the framework aims to draw more attention to the observation that the preferential attachment mechanism is an application of the WRS problem and allow the implementation of the BA model using different random sampling schemes. Finally, the `random-sampling` package during its development has facilitated the experiments performed in the context of this thesis and allowed observations to be made for critical components of the study.

6.1.1 Overview

The `random-sampling` package aims to be a collection of algorithms in Java 8 for the problem of random sampling and weighted random sampling. The library is hosted on Github.com[2] and is published in the Maven Central[3] repository. The library, as of the time of writing of this thesis (v0.18), contains the implementations of various reservoir sampling algorithms, both weighted and unweighted, but other types of

[2]https://github.com/gstamatelat/random-sampling
[3]https://search.maven.org/search?q=gr.james.random-sampling

random sampling methods are planned for the future. For the case of the weighted algorithms, each implementation may interpret the weight parameter in a different way; the exact mechanism at which this parameter is being utilized is documented in the class level notes. For the unweighted algorithms, all implementations are equivalent to each other with respect to the first order inclusion probabilities. The package also contains other independent utilities related to random sampling. For the purposes of the explanations that follow, we consider the population size to be labeled as n and the desired sample size to be labeled as k.

Public API The two main interfaces that dictate the implementation details are `RandomSampling` and `WeightedRandomSampling`; the first allows the implementation of unweighted methods and the second allows the implementation of WRS schemes. Both methods allow the use of generic data types `<T>` within the implementations. The main methods for the `RandomSampling` interface are:

`boolean feed(T item)` Feeds an item to the algorithm. This method simulates the iteration of the elements from the stream and is described in this way to allow the implementation of one pass techniques of sampling. The return value of this method indicates whether there was a change in the underlying reservoir as a result of this call; a value of `true` indicates that `item` was accepted in the reservoir and a value of `false` indicates that `item` was rejected. This method runs in constant time on average.

`Collection<T> sample()` Returns an unmodifiable view of the sample as a generic `Collection` of type `<T>` which has been created from the lifetime of the instance. This method runs in constant time (even with respect to k) as it returns a read-only pointed to the underlying data structure of the reservoir. The unmodifiability is enforced in order to disallow unintentional reservoir manipulation by the API caller.

The `WeightedRandomSampling` interface, which is a super-interface of `RandomSampling` also implements the following method:

`boolean feed(T item, double weight)` Feeds an item to the algorithm using the specified weight as the probability parameter. The interpretation of the value of `weight` is implementation-specific while the semantics of the return value are identical to that of the unweighted version of this method. The unweighted version of this method `feed(T item)` provides a default weight which is also implementation-specific.

Moreover, there are convenience decorate methods around the `feed` method to allow feeding concrete (lists, sets and others) and non-concrete (decorators, implicit iterators and others) collections. The `sample` method is designed to return a `Collection` instead of the more specific classes (for example `List` or `Set`) as the `Collection` class has

only the `size` and `iterator` methods and no other operation should be enforced, for example the elements may not have an order (`List`) or may not be unique (`Set`).

Complexity A fundamental principle of reservoir based sampling algorithms is that the memory complexity is linear with respect to the reservoir size $\mathcal{O}(k)$. Furthermore, the sampling process is performed using a single pass of the stream. The amount of RNG invocations vary among the different implementations.

Duplicates A `RandomSampling` algorithm does not keep track of duplicate elements because that would result in a linear memory complexity. Thus, it is valid to feed the same element multiple times in the same instance. For example it is possible to feed both `x` and `y`, where `x.equals(y)`. The algorithm will treat these items as distinct, even if they are reference-equals (`x == y`). As a result, the final sample `Collection` may contain duplicate elements. Furthermore, elements need not be immutable and the sampling process does not rely on the elements' `hashCode()` and `equals()` methods.

Weights The interpretation of the weight may be different for each `WeightedRandomSampling` implementation. This is inline with the observations and related work previously discussed in Section 5, for example, in Efraimidis (2015) two possible interpretations of the weights are mentioned. As a result, implementations of this interface may not exhibit identical behavior, as opposed to the `RandomSampling` interface. The contract of this interface is, however, that a higher weight value suggests a higher probability for an item to be included in the sample. Implementations may also define certain restrictions on the values of the weight quantity.

Determinism Certain implementations rely on elements of the JRE that are not deterministic, for example `PriorityQueue` and `TreeSet`. The side effect of this is that weighted algorithms are not deterministic either because they typically rely on these data structures. This phenomenon is more prevalent in the presence of ties, where there could be instances of different samples, even with the same seed and the same weighted elements.

Precision Many implementations have an accumulating state which causes the precision of the algorithms to degrade as the stream becomes bigger. An example might be a variable state which strictly increases or decreases as elements are read from the stream. Because the implementations use finite precision data types (usually `double` or `long`), this behavior causes the precision of these implementations to degrade as the stream size increases.

Table 6.1: Algorithms implemented in the `random-sampling` package as of the time of writing of this thesis. The table shows the implementation name as well as the acceptable weight range. In the case of unweighted algorithms, the column of the weights contains a dash.

Algorithm	Implementation	Weights
R by Waterman (Knuth, 1997)	`WatermanSampling`	–
X (Vitter, 1985)	`VitterXSampling`	–
Z (Vitter, 1985)	`VitterZSampling`	–
L (K.-H. Li, 1994b)	`LiLSampling`	–
Chao (1982)	`ChaoSampling`	$[0, \infty)$
A-Res (Efraimidis & Spirakis, 2006)	`EfraimidisSampling`	$[0, \infty)$
Ohlsson (1998)	`SequentialPoissonSampling`	$[0, \infty)$
Rosén (1997a, 1997b)	`ParetoSampling`	$[0, \infty) \setminus 1.0$

Overflow Related to the concept of precision, overflow refers to the situation where the precision has degraded into a non-recurrent state that would prevent the algorithm from behaving consistently. In these cases the implementation will throw `StreamOverflowExceptior` to indicate this state.

6.1.2 List of Implementations

Table 6.1 summarizes the algorithms implemented in the `random-sampling` package as of the time of writing this thesis, including the relevant references of the utilized approaches. The table also displays the acceptable weight range of the implementations in the case of the weighted algorithms or a dash for the case of unweighted approaches.

6.2 The `social-influence` Framework

Concepts of social influence, opinion diffusion or other processes and mechanisms commonly found in social networks play a significant role as both a theoretical and experimental tool for the parts comprising this thesis. In particular, the theoretical background of Chapters 3 and 4 are based on similarity or proximity measures of graph theoretic origin on the Twitter and Foursquare social networks respectively. Furthermore, the experimental setup of both chapters is also based on accurate social network models that typically simulate user behavior in social networks. The social dynamics problems discussed in this thesis are diverse and often might not be related, for example the typical social diffusion models with the MinLA problem, but they have one property in common: they rely on structural analysis and are based on the interpretation of the social network as a graph.

In light of the above and as a natural need for experimentation, the `social-influence` library was developed. It allows the execution of various algorithms under the scope of social networks and social interactions are implemented under the perspective of a social network in a graph data structure. While such frameworks and methods exist, they are usually implemented as part of a different context and are not typically part of a complete platform for the investigation of social interactions. In fact, some of the methods implemented in the `social-influence` framework are not observed under the context of Social Network Analysis, for example the MinLA problem that has been studied in Chapter 3 for the examination of the political orientation of Twitter users.

The `social-influence` framework places these seemingly unrelated methods under a common tool to study social networks and social behavior when social networks are seen under the perspective of a graph structure. It has fully, partially, or in combination with other utilities, accompanied the experiments of Chapters 3, 4 and 5 in this thesis.

6.2.1 Overview of the APIs

The `social-influence` package is a collection of Algorithms in Java 8 to solve various problems of social influence and study social interactions in networks modeled as graphs. The library is hosted on Github.com[4] and is published in the JitPack[5] packaging utility. More information about the types of APIs served via the package are given below.

Public API summary The following list is a summary of the categories of operations contained in the public API.

1. Implementation of the graph data structure (package `graph`)
2. Input and Output (I/O) for graph import and export utilities (package `io`)
3. Implementations of various algorithms (package `algorithms`)
   (a) Distance quantification (package `distance`)
   (b) Graph generators (package `generators`)
   (c) Layout (package `layout`)
   (d) Scoring (package `scoring`)
   (e) Similarity measures (package `similarity`)
   (f) Various other algorithms and tools

[4]https://github.com/gstamatelat/social-influence
[5]https://jitpack.io/#gstamatelat/social-influence

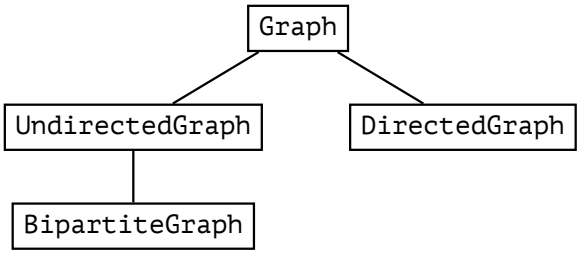


Figure 6.1: Interface hierarchy of the graph implementation package of the `social-influence` framework. The implementations of all interfaces support weighted edges despite the names of the interfaces not conveying this property.

The structure of the framework is such that the concepts of object-oriented programming and the Java programming language are being utilized in order to produce an experience tightly integrated with the Java ecosystem. Related categories of algorithms are indirectly grouped via abstract classes, where possible, and the use of interfaces allows the configurability and extensibility of the framework.

In the following sections, more details are provided for each of the categories of the public API.

6.2.2 Implementation of the graph data structure

The package contains implementations of various types of graphs that are useful for Social Network Analysis, namely undirected graphs, directed graphs, weighted graphs and bipartite graphs. The package is organized in a hierarchy of interfaces, abstract classes, classes and decorator classes. Figure 6.1 shows the interface hierarchy in the framework which classifies the types of graphs in the framework. The highest level interface is the `Graph` which represents a set of vertices without any particular direction of edges defined. The types of vertices and the types of edges are generic so that the full signature of the `Graph` interface is `Graph<V, E>`. The core functionality of the `Graph` interface can be summarized in the following selection of methods:

`Set<V> vertexSet()` Returns an unmodifiable set of the vertices contained in this graph.

`boolean containsVertex(V v)` Checks if the graph contains the specified vertex.

`int vertexCount()` Returns the number of vertices in this graph.

`Iterator<V> iterator()` Get the read-only `Iterator` over the vertices of this graph. This method is also the implementation of the `Iterable` interface.

`boolean addVertex(V v)` Insert the specified vertex `v` to the graph.

`boolean removeVertex(V v)` Removes a vertex from the graph if it is present.

`Graph<V, E> asUnmodifiable()` Returns an unmodifiable decorator around this graph.

`Graph<V, E> toImmutable()` Creates and returns an immutable graph from a copy of this graph. An immutable graph is a copy of the original graph, where the

mutation methods throw `UnsupportedOperationException`.

`Graph<V, E> subGraph(Set<V> vertices)` Returns a subgraph view of this graph that only contains the supplied vertices along with their interconnections. The resulting graph is backed by this graph and will reflect changes to it.

The edge related methods are individually defined in the `UndirectedGraph` and `DirectedGraph` interfaces, which are direct descendants of `Graph` in the interface hierarchy. Specifically, a selection of methods defined in the `UndirectedGraph` interface are the following.

`UndirectedEdge<V, E> findEdge(V v, V w)` Returns the `UndirectedEdge` object representing the edge connecting `v` and `w`, or `null` if there is no such edge. The `UndirectedEdge` interface contains the properties `v()`, `w()`, `weight()` and `value()`, where `v` and `w` represent the ends of the edge, `weight` represents its weight and `value` represents the custom object attached to this edge by the user.

`Set<V> adjacent(V v)` Returns a set of all incident vertices of `v`.

`Iterable<UndirectedEdge<V, E>> edges()` Returns an `Iterable` of all the edges in this graph. The result is lazily populated and does not create a concrete collection that holds the result.

`int degree(V v)` Returns the degree of a vertex. Edge to self is included (if present).

`UndirectedEdge<V, E> addEdge(V v, V w, E edge, double weight)` Creates an edge connecting `v` and `w` with its weight set to `weight`. Also attaches the object `edge` to the edge.

`UndirectedEdge<V, E> removeEdge(V v, V w)` Remove the edge with the specified endpoints `v` and `w`, if it exists.

Similarly, the `DirectedGraph` interface contains the following selection of methods.

`DirectedEdge<V, E> findEdge(V source, V target)` Returns the `DirectedEdge` object representing the edge from `source` to `target`, or `null` if there is no such edge. The `DirectedEdge` interface contains the properties `source()`, `target()`, `weight()` and `value()`, where `source` and `target` represent the directed ends of the edge, `weight` represents its weight and `value` represents the custom object attached to this edge by the user.

`Set<V> adjacentOut(V v)` Returns a set of all incident vertices of `v` that receive an edge from `v`.

`Set<V> adjacentIn(V v)` Returns a set of all incident vertices of `v` that supply an edge to `v`.

`DirectedEdge<V, E> addEdge(V source, V target, E edge, double weight)` Creates an edge connecting `source` and `target` (in this direction) with its weight set to `weight`. Also attaches the object `edge` to the edge.

`DirectedEdge<V, E> removeEdge(V source, V target)` Remove the edge that connects `source` with `target` (in this direction), if it exists.

Finally, the `BipartiteGraph` interface represents a bipartite graph and it extends the `UndirectedGraph` interface with various methods for the manipulation of the two disjoint sets of the vertices, namely `A` and `B`. A selection of the extension methods appears below.

`Set<V> setOf(V v)` Returns the disjoint set of this bipartite graph that contains the specified vertex.

`Set<V> vertexSetA()` Returns the set of vertices comprising the disjoint set `A`.

`Set<V> vertexSetB()` Returns the set of vertices comprising the disjoint set `B`.

`boolean addVertexInA(V v)` Inserts a new vertex into the disjoint set `A`.

`boolean addVertexInB(V v)` Inserts a new vertex into the disjoint set `B`.

`BipartiteGraph<V, E> asSwapped()` Returns a view of this graph, in which the vertex sets `A` and `B` are swapped. This method returns immediately and is backed by the original graph is a lazy way.

The major interfaces defined above have default implementations that can be typically be accessed and constructed using the static `create()` method. All concrete classes have reasonable and consistent implementations of the `toString`, `hashCode` and `equals` methods.

6.2.3 Input/Output

Provides I/O functionality for the importing of graphs from file system or network resources and the export of graphs. The exporting or importing of the objects attached to vertices or edges is performed via a serialization/deserialization technique that transforms the objects into compatible text for each file format. Four major types of formats are supported:

CSV Input and output of graph files as comma-separated values representing the adjacency matrix of the graph. In this file format, the vertex objects cannot be stored.

Edge list Import of export graph files from a text file containing a list of edges that are typically separated by comma. This is also a valid CSV file but different from the CSV format that stores the adjacency matrix. In this format, unconnected vertices cannot be stored as there is no record in the edge list for these nodes.

Dot Import and export graphs based on Graphviz (Ellson et al., 2001) dot format. Only a subset of this file format is supported in order to make exporting to the visualization engine of Graphviz convenient.

JSON Import and export graphs based on a format that resembles the JSON graph specification project[6]. This file format is more flexible than the other formats but is not fully supported as of the time of writing this thesis.

6.2.4 Implementation of algorithms

The algorithms are implemented on separate packages but are not functionally independent as they can have inter-dependencies. Below is a list of the core categories of implemented methods which highlights their structure and public interface.

Distance quantification The package `distance` comprises algorithms for calculation of the distances (shortest paths) among the nodes in a graph. The implementations are categorized based on the type of calculation is involved, namely the all-pairs shortest paths (the distances among all pairs of vertices are computed), the source shortest paths (the distances among a vertex and all other nodes are computed) and the source-sink shortest paths (the distance between only one pair of vertices is computed). The implemented algorithms include Dijkstra's algorithm (Dijkstra et al., 1959) and the Floyd–Warshall algorithm (Floyd, 1962).

Graph generators The public API of the framework distinguishes between graph generators (via the interface `GraphGenerator`) and random graph generators (via the interface `RandomGraphGenerator`). The deterministic graph generators comprise the most typical toy case generators, for example paths, circles, complete graphs, directed paths, grids, wheels and others. Regarding the random graph generators, the framework includes the implementations of the Barabási–Albert model (Barabási & Albert, 1999), the $\mathcal{G}(n, p)$ random graph model (Gilbert, 1959) and the Watts–Strogatz model (Watts & Strogatz, 1998).

Layout The package comprises the implementation of algorithms related to the layout of graphs. First, an algorithm based on the method in Jarvis and Shier (1999) is implemented that determines the period of a directed graph. Because this algorithm only works for strongly connected graphs, the implementation also utilizes Theorem 8.3.1 in M. O. Jackson (2010) which states that

> A graph is convergent if and only if every set of nodes that is strongly connected and closed is aperiodic.

[6] https://github.com/jsongraph/json-graph-specification

The period of the directed graph is calculated as the LCM of the periods of the strongly connected and closed components. The algorithm finds use on certain social phenomena, for example the DeGroot model, where the convergence is based on the period of the given directed graph.

Furthermore, the Minimum Linear Arrangement problem is introduced in this package as it is related to the properties of a graph with respect to its layout. The package contains the implementation of the search algorithm for the solution of the MinLA problem that is described in Section 3.3.3.

The package finally implements the basic formula of the local clustering coefficient, which is also required for the study of several social processes and is typically used as a measure of comparison among social graphs and social graph generators.

Scoring The package contains various methods and utilities that apply a score value on the vertices of a graph. Typical scenarios of this interface are centrality measures, which apply a score value on each vertex that represents its eccentricity in the network, and opinion diffusion models, which apply a score on each vertex that represents its opinion on a particular subject. The implemented centralities include the closeness centrality (Bavelas, 1950) and its closely related decay centrality (M. O. Jackson, 2010, Section 2.2.4) and harmonic centrality (Rochat, 2009) suitable for unconnected graphs, PageRank (Page et al., 1999) and the Hyperlink-Induced Topic Search (HITS) algorithm (Kleinberg et al., 1998). The package also includes the implementation of the DeGroot model of opinion diffusion (Degroot, 1974).

The API distinguishes between the single vertex scoring methods and the multiple vertex scoring methods. Specifically, the single vertex scoring methods are able to compute the score for a single vertex independently of the other scores. In contrast, the multiple vertex scoring methods have to compute more than one score when computing the score of a single vertex. As a result, their difference lies in their operation and not on the public API and is meant to address performance requirements. For example, the implementation of closeness centrality is considered a single vertex operation since one execution of Dijkstra's algorithm is able to produce the score of a vertex. In contrast, the PageRank algorithm is not able to produce a single score without computing the scores of all vertices and, as such, is considered a multiple vertex scoring method. Similarly, the application of a all-pairs shortest paths algorithm for the closeness centrality would result in a multiple vertex scoring method.

Similarity measures The package contains vertex similarity measures, which are functions that return a score value on 2 input vertices. The score is typically a decimal value in the range $[0, 1]$ but some functions return scores in the range $[-1, 1]$ or even an unbounded range. Each function interprets the concept of similarity from a different perspective but there is, in general, positive correlation among them. There are imple-

mentations of popular similarity functions, such as the cosine similarity, the Jaccard index, the overlap coefficient, the Pearson correlation coefficient, the simple matching coefficient (SMC) and the Sørensen–Dice coefficient (F1 score), and others. The set theoretic measures can be easily implemented by considering the compared quantities as the sets of neighbors of the vertices; in this case it is generally acceptable that the more common neighbors two vertices have, the more similar they are. Other measures, such as the Pearson correlation coefficient and the cosine similarity, are implemented by considering each vertex as a binary list of n points where each value represents the existence of absence of an edge to the respective target vertex.

It is worth noting that bipartite graphs have special implementations that may be differentiated from the general implementation for consistency and correctness reasons. For example, the Pearson correlation coefficient is implemented in bipartite graphs such that n is considered the order of the other disjoint set of which the input vertices belong. Since there can be no edge among vertices of the same set, without this special modification the two list would contain zero padding of length equal to the order of the disjoint set of the input vertices and, as a result, the similarity value would be biased towards zero (whether positively or negatively). The same modification applies for other measures for the same reasons, for example in simple matching coefficient where the disjoint set of the input vertices is not included in the calculation of M_{00} (number of vertices that are not connected to either of the input nodes).

Other utilities Several other useful utilities for social network analysis are implemented in the `social-influence` framework. Often described in a clustering setting is the *vertex k-center problem*, which comprises the problem of finding a subset C of k vertices such that

$$\max_{v \in V} \min_{c \in C} d(v, c)$$

is minimized. Here, $d(a, b)$ is the distance among the vertex a and b. The interpretation of the formula is to minimize the maximum distance of any vertex with its closest center, so that no vertex is too far away from a center (a node in C). An abstract greedy algorithm is implemented that attempts a solution of this problem by consecutively selecting the vertices in the C such that the next one is farthest from the current centers. The abstract algorithm accepts the distance function, which needs to satisfy the triangle inequality.

Several connectivity related algorithms are also implemented, for example Kosaraju's algorithm (Sharir, 1981) for the identification of connected components in a directed graph and depth-first and breadth-first search algorithms.

Chapter 7

Conclusions

In this thesis, we investigated the extraction of knowledge from social networks and utilized it in order to provide solutions to common problems that arise in social networks. Our analysis is performed under the assertion that the underlying knowledge of the social network is generated by the users themselves and is based on their behavior. Our assumption, which is demonstrated to be true, is that the social behavior of users is consistent with their beliefs or tastes, a phenomenon that is commonly referred to as *selective exposure*. Our social network analysis methods are purely structural and rely on the social ties among individuals or entities exclusively.

For the applications we examined in this thesis, we utilized knowledge from online social networks (or social media). These social platforms nowadays consist of billions of active users that generate massive volumes of information (structural or non-structural) daily. Our core idea is that users often exhibit public behavior that is consistent with their beliefs or tastes. In our applications, we considered that, on Twitter, users typically select their politically related followers based on their political profile while on Foursquare we considered that users group POIs based on criteria that depend on their own definition of POI similarity, which may be formed due to their past experiences or future plans. While it is common for users to generate such information, sometimes without this realization, users also seek to find information in social networks that is relevant to them, for example news updates, recommendations and others.

In terms of the methodology employed, our main focus is social network analysis. We have utilized common and established methods in the literature in order to imprint a physical property of the network that is relevant to individual applications, as well as novel methods that haven't previously appeared in the context of social network analysis. A common recurrent theme throughout the research presented in this thesis is the concept of graph projections, which is a compression of a –usually bipartite– graph into a unipartite graph by quantifying the pairwise relations among the nodes of one of the disjoint sets. We have shown this type of transformation to be effective in terms of the compression of the network and in terms of the physical interpretation of the

resulting pairwise similarities. Meanwhile, we have applied novel methods, such as the minimum linear arrangement (MinLA) problem, to solve existing, higher abstraction level problems, and to imprint a component of information existing in social networks that was missing in prior literature. Finally, the datasets and primitive data that were acquired for the research are attached as supplementary material in the respective works. We also presented the interface of the software tools and frameworks that were developed during the research.

The first examined application on this research was the identification of political affinity of users in the Twitter network. Twitter was selected due to its profound political semantics which is due to the presence of politically related actors, for example politicians, party representatives, candidates, news media and others. Our goal was to identify the political affinity of certain nodes of interest (NOIs), the Greek MPs and the most popular news media, as these are typically the most politically influential actors in the social network and, hence, the most important to study. Our structural analysis utilizes the follower connections of the NOIs exlusively, in order to establish the consensus about them, rather than what the NOIs' own actions reveal about them. We managed to identify multiple perspectives of political orientation in the multi-party political scene in Greece; first using clustering methods in the follower network to classify the NOIs into their respective political parties and then using the minimum linear arrangement problem to arrange the NOIs into the left-to-right political spectrum axis. Our results were then compared with the approach using the DeGroot model and random walks which shows remarkable consistency with our existing findings. Finally, we apply our methodology on the most popular news media as well to uncover their political leaning, either in relation to each other, or in relation to political parties. Our results are juxtaposed with an online expert survey by a group of political scientists to demonstrate the validity of our approach and the significance of our findinds. Our analysis shows that the information within the dataset that originates from the actions and online behavior of individual anonymous users to follow certain NOIs is very rich and useful and demonstrates the phenomenon of *selective exposure*.

The second application examined in this thesis was relevant to tourism and location recomendations. Here, we used the Foursquare online portal, a location based social network that interconnects users and locations or points of interest. The novelty of our approach is the use of a unique type of interconnected data: the point of interest lists. POI lists are collections of POIs that are generated by users, where each user utilizes their own criteria when synthesizing such list. Our assumption is that users are consistent when placing POIs in lists as, often, these lists can be thought as To-Visit lists since people often want to visit more places than they actually do. As a result, we consider POIs inside a list to be semantically related, with respect to at least one measure, as they are generally consistent with the list creator's interests and tastes. This property allowed us to quantify the pairwise relations among the POIs of the context using various projection methods and quantify the similarities among them

as this is imprinted by the users of the network themselves. These similarities can, in turn, drive a recommendation system based on the principle of "*recommend the most similar places that the user finds interesting*". Experimental results show that our assumptions are reasonable. Our method is applied on a dataset surrounding two areas in northern Greece with significant tourism activity. The evaluation shows interesting observations regarding the relative effectiveness among the projection measures as well as the effectiveness of the system with respect to the baseline popularity measures that are generally difficult to outperform. We also discuss several interesting properties regarding the diversity of the recommendation system, the correlation among the projections and the correlation of the similarities with physical properties of the POIs, such as their location, their categories and their rating. With this analysis, we hope to inspire the use of item lists in the future for either recommendation systems or other applications.

In the last core chapter of this thesis, we presented the development of a new class of algorithms that accurately implement the preferential attachment mechanism in a growing network formation algorithm. The preferential attachment mechanism can explain certain characteristics in real social networks and is the most common way of generating scale-free graphs, i.e. graphs with a degree distribution that follows a power law. We have found that existing models in the literature are only approximations with respect to the proportionality of the inclusion probabilities to the degrees of the vertices. For this reason, we designed an accurate implementation of the preferential attachment mechanism that respects the proportionality of the inclusion probability with the node degrees while at the same time it runs in time proportional to the order of the generated graph. We have analytically confirmed the correctness and the running time of the model and presented computer simulations to demonstrate its properties. Our results demonstrate the tight association between the random sampling problem and the preferential attachment mechanism, and highlight that the fundamental concept of high-order inclusion probabilities when applied to the preferential attachment mechanism can have an impact on the clustering properties of the generated graph.

In this closing part of the thesis, we briefly discuss some general remarks and results from the conducted research and present some interesting directions for future research. Initially, we have observed that our structural analysis methodology has been applied with little preprocessing or –in certain cases– no preprocessing at all. In particular, in our application of the Twitter dataset, we applied the bipartite projections[1] on the primitive data retrieved from the Twitter public API without any preprocessing involved, while on the recommender application of the Foursquare network, we had applied only minimal preprocessing to filter out the lists that contain too few POIs. Our analysis and results indicate that our methods achieve surprisingly high effectiveness even without the presence of preprocessing or data cleansing.

[1]We consider the projections as part of the methodology and not as preprocessing.

This phenomenon might be attributed to the type of primitive data that we have utilized in this research, which can be considered a form of user generated content. The idea behind this type of information is that users, through their actions in the social platform, are synthesizing primitive data that can describe their beliefs, tastes or some other component of their online presence. Through this process, users contribute to a large database of primitive information consisting of structural or non-structural data, without typically making any addditional effort other than using the services provided by the social network service itself. This property could be one of the reasons why preprocessing was largely unnecessary in the applications presented in this thesis. We note that structural data are sometimes part of automatic filtering performed by social services, for example Facebook has a mechanism that disallows certain connection requests, but to our knowledge and with respect to our applications there is no such mechanism imposed by Twitter or Foursquare. Our results indicate that utilizing user generated content in future studies is a promising direction, considering that the amount of such type of information is only expected to grow.

Furthermore, the analyses presented in this research indicated, to a certain extent, that the novel application of graph theoretic concepts in social network analysis can reveal the underlying physical properties of the network. In particular, we showed that the minimum linear arrangement problem can be applied to the Twitter dataset described in Chapter 3 in order to reveal the underlying placement of MPs and news media in the left-to-right political axis. The novel application of this problem was also shown to be quite effective with respect to the placement of individual nodes in the spectrum. As a result, we argue that the application of novel graph theoretic methods should be pursued in the future as an attempt to reveal different perspectives of information from social networks.

Finally, the results presented in Chapter 5 demonstrate the tight relation between the preferential attachment mechanism, a common property in social network analysis, with the random sampling problem. We utilized this connection in order to algorithmically describe a method of a precise and effecient growing random graph generator based on the Barabási–Albert model. We discovered that, despite the preferential attachment mechanism being very common in the literature, its algorithmic analysis is limited and should be pursued in the future. This algorithmic point of view is of particular interest when combined with random sampling, a problem that has seen extensive advancements in terms of efficient computer implementations. Another area of future study in the above scenario is the higher order inclusion probabilities of the growing preferential attachment process that can highlight the probability of a group of nodes to gain common neighbors. Our analysis indicates that the concept of high order probabilities is important in the context of social networks as it involves methodology that is commonly used for social network analysis, such as the projection methods that were described throughout this research that convey the similarities among the nodes of the network, often with respect to their common neighbors.

Bibliography

Yule, G. U. (1925). II.—A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F. R. S. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character*, *213*(402-410), 21–87. https://doi.org/10.1098/rstb.1925.0002

Bavelas, A. (1950). Communication patterns in task-oriented groups. *The journal of the acoustical society of America*, *22*(6), 725–730.

Goodman, R., & Kish, L. (1950). Controlled selection—a technique in probability sampling. *Journal of the American Statistical Association*, *45*(251), 350–372. https://doi.org/10.1080/01621459.1950.10501130

Yates, F., & Grundy, P. M. (1953). Selection Without Replacement from Within Strata with Probability Proportional to Size. *Journal of the Royal Statistical Society: Series B (Methodological)*, *15*(2), 253–261. https://doi.org/10.1111/j.2517-6161.1953.tb00140.x

Lazarsfeld, P. F., Merton, R. K., et al. (1954). Friendship as a social process: A substantive and methodological analysis. *Freedom and control in modern society*, *18*(1), 18–66.

Dijkstra, E. W., et al. (1959). A note on two problems in connexion with graphs. *Numerische mathematik*, *1*(1), 269–271.

Erdős, P., & Rényi, A. (1959). On random graphs i. *Publicationes Mathematicae (Debrecen)*, *6*, 290–297.

Gilbert, E. N. (1959). Random graphs. *The Annals of Mathematical Statistics*, *30*(4), 1141–1144.

Floyd, R. W. (1962). Algorithm 97: Shortest path. *Communications of the ACM*, *5*(6), 345.

Price, D. J. (1963). De solla. *Little science, big science*, 30–31.

Hajek, J. (1964). Asymptotic Theory of Rejective Sampling with Varying Probabilities from a Finite Population. *The Annals of Mathematical Statistics*, *35*(4), 1491–1523. https://doi.org/10.1214/aoms/1177700375

Jessen, R. J. (1969). Some methods of probability non-replacement sampling. *Journal of the American Statistical Association*, *64*(325), 175–193. https://doi.org/10.1080/01621459.1969.10500962

Degroot, M. H. (1974). Reaching a consensus. *Journal of the American Statistical Association*, *69*(345), 118–121. https://doi.org/10.1080/01621459.1974.10480137

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 35–41.

Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of anthropological research*, *33*(4), 452–473.

Hanif, M., & Brewer, K. R. W. (1980). Sampling with Unequal Probabilities without Replacement: A Review. *International Statistical Review / Revue Internationale de Statistique*, *48*(3), 317. https://doi.org/10.2307/1402944

Bruijn, N. G. d. (1981). *Asymptotic methods in analysis* (Dover). Dover Publications.

Sharir, M. (1981). A strong-connectivity algorithm and its applications in data flow analysis. *Computers & Mathematics with Applications*, *7*(1), 67–72.

Chao, M. T. (1982). A general purpose unequal probability sampling plan. *Biometrika*, *69*(3), 653–656. https://doi.org/10.1093/biomet/69.3.653

Brewer, K. R., & Hanif, M. (1983). *Sampling with unequal probabilities* (Vol. 15). Springer Science & Business Media.

Eades, P. (1984). A heuristic for graph drawing. *Congressus numerantium*, *42*, 149–160.

Vitter, J. S. (1985). Random sampling with a reservoir. *ACM Trans. Math. Softw.*, *11*(1), 37–57. https://doi.org/10.1145/3147.3165

Action, R. (1993). The rise of the medici. *American Journal of Sociology*, *98*, 1259–1319.

Li, K.-H. (1994a). A computer implementation of the yates-grundy draw by draw procedure. *Journal of Statistical Computation and Simulation*, *50*(3-4), 147–151. https://doi.org/10.1080/00949659408811606

Li, K.-H. (1994b). Reservoir-sampling algorithms of time complexity $\mathcal{O}(n(1+log(N/n)))$. *ACM Trans. Math. Softw.*, *20*(4), 481–493. https://doi.org/10.1145/198429.198435

Doğrusöz, U., Madden, B., & Madden, P. (1997). Circular layout in the graph layout toolkit. In S. North (Ed.), *Graph drawing* (pp. 92–100). Springer Berlin Heidelberg.

Knuth, D. E. (1997). *The art of computer programming, volume 2 (3rd ed.): Seminumerical algorithms*. Addison-Wesley Longman Publishing Co., Inc.

Resnick, P., & Varian, H. R. (1997). Recommender systems. *Communications of the ACM*, *40*(3), 56–58.

Rosén, B. (1997a). Asymptotic theory for order sampling. *Journal of Statistical Planning and Inference*, *62*(2), 135–158. https://doi.org/10.1016/S0378-3758(96)00185-1

Rosén, B. (1997b). On sampling with probability proportional to size. *Journal of Statistical Planning and Inference*, *62*(2), 159–191. https://doi.org/10.1016/S0378-3758(96)00186-3

Kleinberg, J. M., et al. (1998). Authoritative sources in a hyperlinked environment. *SODA*, *98*, 668–677.

Ohlsson, E. (1998). Sequential poisson sampling. *Journal of official Statistics*, *14*(2), 149.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440–442. https://doi.org/10.1038/30918

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *science*, *286*(5439), 509–512.

Jarvis, J., & Shier, D. R. (1999). Graph-theoretic analysis of finite markov chains. *Applied mathematical modeling: a multidisciplinary approach*, 85.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The pagerank citation ranking: Bringing order to the web.* (tech. rep.). Stanford InfoLab.

Bollobás, B., Riordan, O., Spencer, J., & Tusnády, G. (2001). The degree sequence of a scale-free random graph process: Degree Sequence of a Random Graph. *Random Structures & Algorithms*, *18*(3), 279–290. https://doi.org/10.1002/rsa.1009

Ellson, J., Gansner, E., Koutsofios, L., North, S. C., & Woodhull, G. (2001). Graphviz—open source graph drawing tools. *International Symposium on Graph Drawing*, 483–484.

Pigott, T. D. (2001). A review of methods for missing data. *Educational Research and Evaluation*, *7*(4), 353–383. https://doi.org/10.1076/edre.7.4.353.8937

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. *Proceedings of the 10th International Conference on World Wide Web*, 285–295. https://doi.org/10.1145/371920.372071

Smyth, B., & McClave, P. (2001). Similarity vs. diversity. In D. W. Aha & I. Watson (Eds.), *Case-based reasoning research and development* (pp. 347–361). Springer Berlin Heidelberg.

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74*(1), 47–97. https://doi.org/10.1103/RevModPhys.74.47

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, *99*(12), 7821–7826.

Holme, P., & Kim, B. J. (2002). Growing scale-free networks with tunable clustering. *Physical Review E*, *65*(2), 026107. https://doi.org/10.1103/PhysRevE.65.026107

Jeh, G., & Widom, J. (2002). Simrank: A measure of structural-context similarity. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 538–543.

Bornholdt, S., & Schuster, H. G. (Eds.). (2003). *Handbook of graphs and networks: From the genome to the internet* (1st ed) [OCLC: ocm50056112]. Wiley-VCH.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, *45*(2), 167–256.

Newman, M. E., & Park, J. (2003). Why social networks are different from other types of networks. *Physical review E*, *68*(3), 036122.

Petit, J. (2003). Experiments on the minimum linear arrangement problem. *Journal of Experimental Algorithmics*, *8*, 1–29. https://doi.org/10.1145/996546.996554

Batagelj, V., & Brandes, U. (2005). Efficient generation of large random networks. *Physical Review E*, *71*(3), 036113. https://doi.org/10.1103/PhysRevE.71.036113

Catanzaro, M., Boguñá, M., & Pastor-Satorras, R. (2005). Generation of uncorrelated random scale-free networks. *Physical Review E*, *71*(2), 027103. https://doi.org/10.1103/PhysRevE.71.027103

Koren, Y. (2005). Drawing graphs by eigenvectors: Theory and practice. *Computers & Mathematics with Applications*, *49*(11-12), 1867–1888. https://doi.org/10.1016/j.camwa.2004.08.015

Newman, M. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, *46*(5), 323–351. https://doi.org/10.1080/00107510500052444

Csardi, G., Nepusz, T., et al. (2006). The igraph software package for complex network research. *InterJournal, complex systems*, *1695*(5), 1–9.

Efraimidis, P. S., & Spirakis, P. G. (2006). Weighted random sampling with a reservoir. *Information Processing Letters*, *97*(5), 181–185. https://doi.org/https://doi.org/10.1016/j.ipl.2005.11.003

Safro, I., Ron, D., & Brandt, A. (2006). Graph minimum linear arrangement by multilevel weighted edge contractions. *Journal of Algorithms*, *60*(1), 24–41. https://doi.org/10.1016/j.jalgor.2004.10.004

Tillé, Y. (2006). *Sampling algorithms*. Springer.

Ambuhl, C., Mastrolilli, M., & Svensson, O. (2007). Inapproximability results for sparsest cut, optimal linear arrangement, and precedence constrained scheduling. *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS '07)*, 329–337. https://doi.org/10.1109/FOCS.2007.40

Feige, U., & Lee, J. R. (2007). An improved approximation ratio for the minimum linear arrangement problem. *Information Processing Letters*, *101*(1), 26–29. https://doi.org/10.1016/j.ipl.2006.07.009

Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, *58*(7), 1019–1031. https://doi.org/10.1002/asi.20591

Schaeffer, S. E. (2007). Graph clustering. *Computer science review*, *1*(1), 27–64.

Zhou, T., Ren, J., Medo, M. š., & Zhang, Y.-C. (2007). Bipartite network projection and personal recommendation. *Physical Review E*, *76*, 046115. https://doi.org/10.1103/PhysRevE.76.046115

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Hagberg, A., Swart, P., & S Chult, D. (2008). *Exploring network structure, dynamics, and function using networkx* (tech. rep.). Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. *Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media (ICWSM '09)*, 361–362.

Berger, Y. G., & Tillé, Y. (2009). Sampling with unequal probabilities. In *Handbook of statistics* (pp. 39–54). Elsevier.

Chierichetti, F., Kumar, R., Lattanzi, S., Mitzenmacher, M., Panconesi, A., & Raghavan, P. (2009). On compressing social networks. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 219–228. https://doi.org/10.1145/1557019.1557049

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, *51*(4), 661–703. https://doi.org/10.1137/070710111

Garrett, R. K. (2009). Politically motivated reinforcement seeking: Reframing the selective exposure debate. *Journal of Communication*, *59*(4), 676–699. https://doi.org/10.1111/j.1460-2466.2009.01452.x

Lancichinetti, A., & Fortunato, S. (2009). Community detection algorithms: A comparative analysis. *Physical Review E*, *80*, 056117. https://doi.org/10.1103/PhysRevE.80.056117

Lattanzi, S., & Sivakumar, D. (2009). Affiliation networks. *Proceedings of the forty-first annual ACM symposium on Theory of computing*, 427–434.

Noack, A. (2009). Modularity clustering is force-directed layout. *Physical Review E*, *79*, 026102. https://doi.org/10.1103/PhysRevE.79.026102

Rochat, Y. (2009). *Closeness centrality extended to unconnected graphs: The harmonic centrality index* (tech. rep.).

Zhou, T., Lü, L., & Zhang, Y.-C. (2009). Predicting missing links via local information. *The European Physical Journal B*, *71*(4), 623–630. https://doi.org/10.1140/epjb/e2009-00335-8

Agresti, A. (2010). *Analysis of ordinal categorical data* (Second). John Wiley & Sons.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*(3), 75–174. https://doi.org/10.1016/j.physrep.2009.11.002

Gentzkow, M., & Shapiro, J. M. (2010). What drives media slant? Evidence from U.S. daily newspapers. *Econometrica*, *78*(1), 35–71. https://doi.org/10.3982/ECTA7195

Goel, S., Broder, A., Gabrilovich, E., & Pang, B. (2010). Anatomy of the long tail: Ordinary people with extraordinary tastes. *Proceedings of the third ACM International Conference on Web Search and Data Mining*, 201–210. https://doi.org/10.1145/1718487.1718513

Golub, B., & Jackson, M. O. (2010). Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, *2*(1), 112–49.

Grafström, A. (2010). *On unequal probability sampling designs* (Doctoral dissertation) [OCLC: 648090592]. Umeå University. Umeå University, Department of Mathematics; Mathematical Statistics, Umeå University. Retrieved December 15, 2020, from http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-33701%20urn:nbn:se:umu:diva-33701

Jackson, L. S., & Forster, P. M. (2010). An empirical study of geographic and seasonal variations in diurnal temperature range. *Journal of Climate*, *23*(12), 3205–3221. https://doi.org/10.1175/2010JCLI3215.1

Jackson, M. O. (2010). *Social and economic networks*. Princeton university press.

Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., & Makse, H. A. (2010). Identification of influential spreaders in complex networks. *Nature physics*, *6*(11), 888–893.

Li, C., Han, J., He, G., Jin, X., Sun, Y., Yu, Y., & Wu, T. (2010). Fast computation of SimRank for static and dynamic information networks. *13th International Conference on Extending Database Technology - EDBT '10*, 465. https://doi.org/10.1145/1739041.1739098

Newman, M. E. J. (2010). *Networks: An introduction* [OCLC: 670040773]. Oxford University Press.

Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, I. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM '10)*, 178–185.

Conover, M. D., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., & Flammini, A. (2011). Political polarization on twitter. *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 89–96.

Kwiatkowska, M., Norman, G., & Parker, D. (2011). PRISM 4.0: Verification of probabilistic real-time systems. In G. Gopalakrishnan & S. Qadeer (Eds.), *Computer aided verification* (pp. 585–591). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-22110-1_47

Pennacchiotti, M., & Popescu, A.-M. (2011). A machine learning approach to Twitter user classification. *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM '11)*, 281–288.

Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to recommender systems handbook. In *Recommender systems handbook* (pp. 1–35). Springer.

Zheng, Y., & Zhou, X. (Eds.). (2011a). *Computing with spatial trajectories*. Springer. https://doi.org/10.1007/978-1-4614-1629-6

Zheng, Y., & Zhou, X. (2011b). *Computing with spatial trajectories*. Springer Science & Business Media.

An, J., Cha, M., Gummadi, K., Crowcroft, J., & Quercia, D. (2012). Visualizing media bias through Twitter. *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM '12), Workshop on the Potential of Social Media Tools and Data for Journalists*, 2–5.

Colomer-de-Simon, P., & Boguñá, M. (2012). Clustering of random scale-free networks. *Physical Review E*, *86*(2), 026120. https://doi.org/10.1103/PhysRevE.86.026120

Hariri, N., Mobasher, B., & Burke, R. (2012). Context-aware music recommendation based on latenttopic sequential patterns. *Proceedings of the sixth ACM conference on Recommender systems - RecSys '12*, 131. https://doi.org/10.1145/2365952.2365979

Kobourov, S. G. (2012). Spring embedders and force directed graph drawing algorithms. *arXiv preprint arXiv:1201.3011*.

Maynard, D., & Funk, A. (2012). Automatic detection of political opinions in tweets. In R. García-Castro, D. Fensel, & G. Antoniou (Eds.), *The semantic web: Eswc 2011 workshops* (pp. 88–99). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-25953-1_8

Noulas, A., Scellato, S., Lathia, N., & Mascolo, C. (2012). A random walk around the city: New venue recommendation in location-based social networks. *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, 144–153. https://doi.org/10.1109/SocialCom-PASSAT.2012.70

Papadopoulos, S., Kompatsiaris, Y., Vakali, A., & Spyridonos, P. (2012). Community detection in social media. *Data Mining and Knowledge Discovery*, *24*(3), 515–554.

Raghavendra, P., Steurer, D., & Tulsiani, M. (2012). Reductions between expansion problems. *Proceedings of the IEEE 27th Conference on Computational Complexity*, 64–73. https://doi.org/10.1109/CCC.2012.43

Verweij, P. (2012). Twitter links between politicians and journalists. *Journalism Practice*, *6*(5-6), 680–691. https://doi.org/10.1080/17512786.2012.667272

Boutet, A., Kim, H., & Yoneki, E. (2013). What's in Twitter, I know what parties are popular and who you are supporting now! *Social Network Analysis and Mining*, *3*(4), 1379–1391. https://doi.org/10.1007/s13278-013-0120-1

Ghaderi, J., & Srikant, R. (2013). Opinion dynamics in social networks: A local interaction game with stubborn agents. *American Control Conference (ACC '13)*, 1982–1987. https://doi.org/10.1109/ACC.2013.6580126

Hannak, A., Sapiezynski, P., Molavi Kakhki, A., Krishnamurthy, B., Lazer, D., Mislove, A., & Wilson, C. (2013). Measuring personalization of web search. *Proceedings of the 22nd International Conference on World Wide Web (WWW '13)*, 527–538. https://doi.org/10.1145/2488388.2488435

Preoţiuc-Pietro, D., Cranshaw, J., & Yano, T. (2013). Exploring venue-based city-to-city similarity measures. *2nd ACM SIGKDD International Workshop on Urban Computing - UrbComp '13*, 1. https://doi.org/10.1145/2505821.2505832

Wang, H., Terrovitis, M., & Mamoulis, N. (2013). Location recommendation in location-based social networks using user check-in data. *21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - SIGSPATIAL'13*, 374–383. https://doi.org/10.1145/2525314.2525357

Berger, N., Borgs, C., Chayes, J. T., & Saberi, A. (2014). Asymptotic behavior and distributional limits of preferential attachment graphs. *The Annals of Probability*, *42*(1). https://doi.org/10.1214/12-AOP755

Borgatti, S. P., & Halgin, D. S. (2014). Analyzing affiliation networks. In J. Scott & P. Carrington (Eds.), *The sage handbook of social network analysis* (pp. 417–433). SAGE Publications Ltd. https://doi.org/10.4135/9781446294413

Golbeck, J., & Hansen, D. (2014). A method for computing political preference among Twitter followers. *Social Networks*, *36*, 177–184. https://doi.org/10.1016/j.socnet.2013.07.004

Munroe, R. (2014). *What if?: Serious scientific answers to absurd hypothetical questions*. Houghton Mifflin Harcourt.

Parmelee, J. H. (2014). The agenda-building function of political tweets. *New Media & Society*, *16*(3), 434–450.

Bao, J., Zheng, Y., Wilkie, D., & Mokbel, M. (2015). Recommendations in location-based social networks: A survey. *GeoInformatica*, *19*(3), 525–565. https://doi.org/10.1007/s10707-014-0220-8

Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political Analysis*, *23*(1), 76–91. https://doi.org/10.1093/pan/mpu011

Efraimidis, P. S. (2015). Weighted random sampling over data streams. In C. Zaroliagis, G. Pantziou, & S. Kontogiannis (Eds.), *Algorithms, probability, networks, and games: Scientific papers and essays dedicated to paul g. spirakis on the occasion of his 60th birthday* (pp. 183–195). Springer International Publishing. https://doi.org/10.1007/978-3-319-24024-4_12

Karagiannis, I., Arampatzis, A., Efraimidis, P. S., & Stamatelatos, G. (2015). Social network analysis of public lists of POIs. *19th Panhellenic Conference on Informatics - PCI '15*, 61–62. https://doi.org/10.1145/2801948.2802031

Lu, W., & Stepchenkova, S. (2015). User-generated content as a research mode in tourism and hospitality applications: Topics, methods, and software. *Journal of Hospitality Marketing & Management*, *24*(2), 119–154. https://doi.org/10.1080/19368623.2014.907758

Schall, D. (2015). Link prediction for directed graphs. In *Social network-based recommender systems* (pp. 7–31). Springer. https://doi.org/10.1007/978-3-319-22735-1_2

Song, Q., Cheng, J., Yuan, T., & Lu, H. (2015). Personalized recommendation meets your next favorite. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 1775–1778. https://doi.org/10.1145/2806416.2806598

Wu, D., Mamoulis, N., & Shi, J. (2015). Clustering in geo-social networks. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, *38*, 47–57.

Alzahrani, T., & Horadam, K. J. (2016). Community detection in bipartite networks: Algorithms and case studies. In J. Lü, X. Yu, G. Chen, & W. Yu (Eds.), *Complex systems and networks: Dynamics, controls and applications* (pp. 25–50). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-47824-0_2

Dredze, M., Osborne, M., & Kambadur, P. (2016). Geolocation for twitter: Timing matters. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1064–1069. https://doi.org/10.18653/v1/N16-1122

Giachanou, A., & Crestani, F. (2016). Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys*, *49*, 1–41. https://doi.org/10.1145/2938640

Grover, A., & Leskovec, J. (2016). Node2vec: Scalable Feature Learning for Networks. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 855–864. https://doi.org/10.1145/2939672.2939754

Hadian, A., Nobari, S., Minaei-Bidgoli, B., & Qu, Q. (2016). ROLL: Fast In-Memory Generation of Gigantic Scale-free Networks. *Proceedings of the 2016 International Conference on Management of Data - SIGMOD '16*, 1829–1842. https://doi.org/10.1145/2882903.2882964

Hashemi, S. H., Clarke, C. L., Kamps, J., Kiseleva, J., & Voorhees, E. M. (2016). Overview of the TREC 2016 contextual suggestion track. *25th Text Retrieval Conference - TREC '16*, 1–10.

Hidasi, B., Quadrana, M., Karatzoglou, A., & Tikk, D. (2016). Parallel Recurrent Neural Network Architectures for Feature-rich Session-based Recommendations. *Proceedings of the 10th ACM Conference on Recommender Systems*, 241–248. https://doi.org/10.1145/2959100.2959167

King, A. S., Orlando, F. J., & Sparks, D. B. (2016). Ideological extremity and success in primary elections: Drawing inferences from the twitter network. *Social Science Computer Review*, *34*(4), 395–415. https://doi.org/10.1177/0894439315595483

Ravi, L., & Vairavasundaram, S. (2016). A collaborative location based travel recommendation system through enhanced rating prediction for the group of users. *Computational Intelligence and Neuroscience*, *2016*, 1–28. https://doi.org/10.1155/2016/1291358

Riquelme, F., & González-Cantergiani, P. (2016). Measuring user influence on twitter: A survey. *Information Processing & Management*, *52*(5), 949–975. https://doi.org/10.1016/j.ipm.2016.04.003

Van Der Hofstad, R. (2016). *Random graphs and complex networks* (Vol. 1). Cambridge university press.

Vijaymeena, M., & Kavitha, K. (2016). A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal*, *3*(1), 19–28. https://doi.org/10.5121/mlaij.2016.3103

Abdollahpouri, H., Burke, R., & Mobasher, B. (2017). Controlling popularity bias in learning-to-rank recommendation. *Proceedings of the Eleventh ACM Conference on Recommender Systems*, 42–46. https://doi.org/10.1145/3109859.3109912

Ahmedi, L., Rrmoku, K., Sylejmani, K., & Shabani, D. (2017). A bimodal social network analysis to recommend points of interest to tourists. *Social Network Analysis and Mining*, *7*(1), 14. https://doi.org/10.1007/s13278-017-0431-8

Arnab, R. (2017). *Survey sampling theory and applications* [OCLC: ocn959875086]. Elsevier/AP, Academic Press, an imprint of Elsevier.

Cenamor, I., de la Rosa, T., Núñez, S., & Borrajo, D. (2017). Planning for tourism routes using social networks. *Expert Systems with Applications*, *69*, 1–9. https://doi.org/10.1016/j.eswa.2016.10.030

Kaminskas, M., & Bridge, D. (2017). Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender

Systems. *ACM Transactions on Interactive Intelligent Systems*, *7*(1), 1–42. https://doi.org/10.1145/2926720

Le, H., Shafiq, Z., & Srinivasan, P. (2017). Scalable news slant measurement using Twitter. *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM '17)*, 584–587.

Marozzo, F., & Bessi, A. (2017). Analyzing polarization of social media users and news sites during political campaigns. *Social Network Analysis and Mining*, *8*(1), 1. https://doi.org/10.1007/s13278-017-0479-5

Rakesh, V., Jadhav, N., Kotov, A., & Reddy, C. K. (2017). Probabilistic social sequential model for tour recommendation. *10th ACM International Conference on Web Search and Data Mining - WSDM '17*, 631–640. https://doi.org/10.1145/3018661.3018711

Wörndl, W., Hefele, A., & Herzog, D. (2017). Recommending a sequence of interesting places for tourist trips. *Information Technology & Tourism*, *17*(1), 31–54. https://doi.org/10.1007/s40558-017-0076-5

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, *18*(3), 91–93. https://doi.org/10.1016/j.tjem.2018.08.001

Arentze, T., Kemperman, A., & Aksenov, P. (2018). Estimating a latent-class user model for travel recommender systems. *Information Technology & Tourism*, *19*(1-4), 61–82. https://doi.org/10.1007/s40558-018-0105-z

Briola, H., Drosatos, G., Stamatelatos, G., Gyftopoulos, S., & Efraimidis, P. S. (2018). Privacy leakages about political beliefs through analysis of twitter followers. *Proceedings of the 22nd Pan-Hellenic Conference on Informatics (PCI '18)*, 16–21. https://doi.org/10.1145/3291533.3291557

Celik, M., & Dokuz, A. S. (2018). Discovering socially similar users in social media datasets based on their socially important locations. *Information Processing & Management*, *54*(6), 1154–1168. https://doi.org/10.1016/j.ipm.2018.08.004

Christoforidis, G., Kefalas, P., Papadopoulos, A., & Manolopoulos, Y. (2018). Recommendation of points-of-interest using graph embeddings. *5th International Conference on Data Science and Advanced Analytics - DSAA '18)*, 31–40. https://doi.org/10.1109/DSAA.2018.00013

David-Negre, T., Almedida-Santana, A., Hernández, J. M., & Moreno-Gil, S. (2018). Understanding European tourists' use of e-tourism platforms. Analysis of networks. *Information Technology & Tourism*, *20*(1-4), 131–152. https://doi.org/10.1007/s40558-018-0113-z

Eirinaki, M., Gao, J., Varlamis, I., & Tserpes, K. (2018). Recommender systems for large-scale social networks: A review of challenges and solutions. *Future Generation Computer Systems*, *78*, 413–418. https://doi.org/10.1016/j.future.2017.09.015

Goyal, P., & Ferrara, E. (2018). Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems*, *151*, 78–94. https://doi.org/10.1016/j.knosys.2018.03.022

Hasan, M., Orgun, M. A., & Schwitter, R. (2018). A survey on real-time event detection from the twitter data stream. *Journal of Information Science*, *44*(4), 443–463. https://doi.org/10.1177/0165551517698564

Karataş, A., & Şahin, S. (2018). Application areas of community detection: A review. *2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT)*, 65–70. https://doi.org/10.1109/IBIGDELFT.2018.8625349

Kefalas, P., Symeonidis, P., & Manolopoulos, Y. (2018). Recommendations based on a heterogeneous spatio-temporal social network. *World Wide Web*, *21*(2), 345–371. https://doi.org/10.1007/s11280-017-0454-0

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology* (Fourth). Sage Publications, Inc.

Ludewig, M., & Jannach, D. (2018). Evaluation of session-based recommendation algorithms. *User Modeling and User-Adapted Interaction*, *28*(4-5), 331–390. https://doi.org/10.1007/s11257-018-9209-6

Newman, M. (2018). *Networks*. Oxford university press.

Nisar, T. M., & Yeung, M. (2018). Twitter as a tool for forecasting stock market movements: A short-window event study. *The Journal of Finance and Data Science*, *4*(2), 101–119. https://doi.org/10.1016/j.jfds.2017.11.002

Quadrana, M., Cremonesi, P., & Jannach, D. (2018). Sequence-aware recommender systems. *ACM Comput. Surv.*, *51*(4). https://doi.org/10.1145/3190616

Ribeiro, F. N., Henrique, L., Benevenuto, F., Chakraborty, A., Kulshrestha, J., Babaei, M., & Gummadi, K. P. (2018). Media bias monitor: Quantifying biases of social media news outlets at large-scale. *Proceedings of the 12th International AAAI Conference on Web and Social Media (ICWSM '18)*, 290–299.

Stamatelatos, G., Gyftopoulos, S., Drosatos, G., & Efraimidis, P. S. (2018). Deriving the political affinity of twitter users from their followers. *2018 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications (ISPA/IUCC/BDCloud/SocialCom/SustainCom)*, 1175–1182.

Tsakalidis, A., Aletras, N., Cristea, A. I., & Liakata, M. (2018). Nowcasting the stance of social media users in a sudden vote: The case of the greek referendum. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 367–376. https://doi.org/10.1145/3269206.3271783

Van der Zee, E., & Bertocchi, D. (2018). Finding patterns in urban tourist behaviour: A social network analysis approach based on TripAdvisor reviews. *Information Technology & Tourism*, *20*(1-4), 153–180. https://doi.org/10.1007/s40558-018-0128-5

Zarrinkalam, F., Kahani, M., & Bagheri, E. (2018). Mining user interests over active topics on social networks. *Information Processing & Management*, *54*(2), 339–357. https://doi.org/10.1016/j.ipm.2017.12.003

Armandpour, M., Ding, P., Huang, J., & Hu, X. (2019). Robust Negative Sampling for Network Embedding. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*, 3191–3198. https://doi.org/10.1609/aaai.v33i01.33013191

Belletti, F., Lakshmanan, K., Krichene, W., Chen, Y.-F., & Anderson, J. (2019). Scalable realistic recommendation datasets through fractal expansions.

Broido, A. D., & Clauset, A. (2019). Scale-free networks are rare. *Nature Communications*, *10*(1), 1017. https://doi.org/10.1038/s41467-019-08746-5

Li, L., Lee, K. Y., & Yang, S.-B. (2019). Exploring the effect of heuristic factors on the popularity of user-curated 'best places to visit' recommendations in an online travel community. *Information Processing & Management*, *56*(4), 1391–1408. https://doi.org/10.1016/j.ipm.2018.03.009

Mighan, S. N., Kahani, M., & Pourgholamali, F. (2019). POI recommendation based on heterogeneous graph embedding. *9th International Conference on Computer and Knowledge Engineering - ICCKE '19*, 188–193. https://doi.org/10.1109/ICCKE48569.2019.8964762

Poulakidakos, S., & Veneti, A. (2019). Political communication and twitter in greece: Jumps on the bandwagon or an enhancement of the political dialogue? In *Censorship, surveillance, and privacy: Concepts, methodologies, tools, and applications* (pp. 1125–1152). IGI Global.

Sainudiin, R., Yogeeswaran, K., Nash, K., & Sahioun, R. (2019). Characterizing the twitter network of prominent politicians and splc-defined hate groups in the 2016 us presidential election. *Social Network Analysis and Mining*, *9*(1), 34. https://doi.org/10.1007/s13278-019-0567-9

Sertkan, M., Neidhardt, J., & Werthner, H. (2019). What is the "Personality" of a tourism destination? *Information Technology & Tourism*, *21*(1), 105–133. https://doi.org/10.1007/s40558-018-0135-6

Gyftopoulos, S., Drosatos, G., Stamatelatos, G., & Efraimidis, P. S. (2020). A twitter-based approach of news media impartiality in multipartite political scenes. *Social Network Analysis and Mining*, *10*, 1–16.

Stamatelatos, G., Gyftopoulos, S., Drosatos, G., & Efraimidis, P. S. (2020). Revealing the political affinity of online entities through their twitter followers. *Information Processing & Management*, *57*(2), 102172. https://doi.org/10.1016/j.ipm.2019.102172

Stamatelatos, G., Drosatos, G., Gyftopoulos, S., Briola, H., & Efraimidis, P. S. (2021). Point-of-interest lists and their potential in recommendation systems. *Information Technology & Tourism*, *23*(2), 209–239.

Stamatelatos, G., & Efraimidis, P. S. (2021a). About weighted random sampling in preferential attachment models. *arXiv preprint arXiv:2102.08173*.

Stamatelatos, G., & Efraimidis, P. S. (2021b). Whole sampling generation of scale-free graphs. *arXiv preprint arXiv:2110.00287*.