Algorithmic Analysis of User Behavior in Social Media

Applications of the knowledge obtained through the behavior of OSN users via structural analysis

Giorgos Stamatelatos

Advisor: Pavlos S. Efraimidis

Dept. Electrical & Computer Engineering,
Democritus University of Thrace

January 25, 2022

Overview & Findings

Part I Political Affinity on Twitter. The behavior of Twitter followers in respect to their beliefs and the effects of their actions in the network.

- Twitter followers can reveal the political orientation of the users they opt to follow.
- Two perspectives of political orientation: categorical (parties) and arrangement (axis).

Part II User Generated POI Lists. What is the criteria that social network users employ to group POIs in lists and how does it help recommender systems.

- POI lists are collections of semantically related POIs.
- The relations can be utilized to drive a recommendation system.

Part III Preferential Attachment. The mechanism under which users select connections in social networks.

- Preferential attachment is an application of *weighted random sampling*.
- Accurate and efficient implementation of the the Barabási–Albert model.

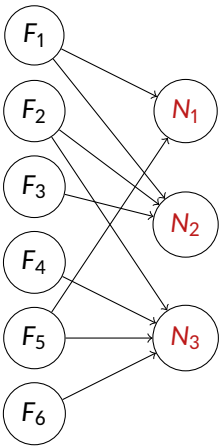
Part I: Political Affinity on Twitter

The behavior of Twitter followers in respect to their beliefs and the effects of their actions in the network.

Overview

- Social Network Analysis (SNA) methods to derive the political affinity of particular Twitter users.
  - Nodes of Interest (NOIs): MPs of the Greek Parliament & Greek news media.
  - Structure-based methods that rely only on the follower sets of the NOIs.
- Two suitable methods: Modularity Clustering, MinLA problem.
  - Partitioned the NOIs in their respective political parties.
  - Arrangement of the political parties in the left-to-right political spectrum.
- Twitter users can reveal valuable political information about the NOIs they opt to follow (*selective exposure*).

Bipartite Dataset & Projections



Primitive Data

NOIs	Followers	Edges
162	750,537	2,492,237

- Twitter snapshot: April 2018.
- Follower relationships as the only source of information.
- The followers *expose* the affinity of the NOIs.

Bipartite Projection

Convert to uni-partite, weighted, complete graph
$w_{ij} = \text{sim}(N_i, N_j)$.

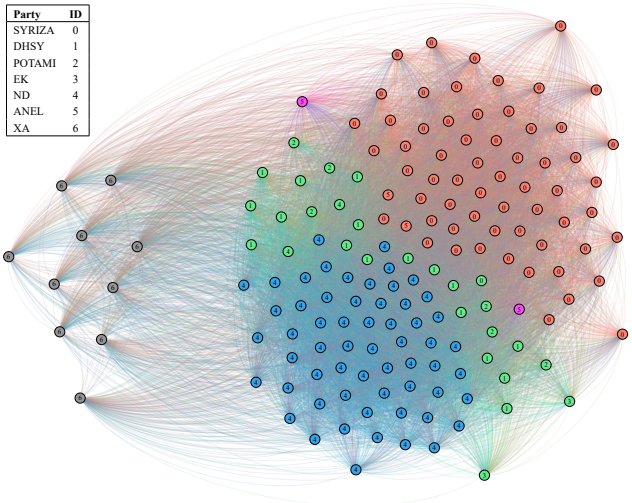
Modularity Clustering Application

- Modularity clustering via Louvain optimization method.
- Clustering partitions the NOIs in their respective parties.
- The overlap coefficient is the best for this setting:

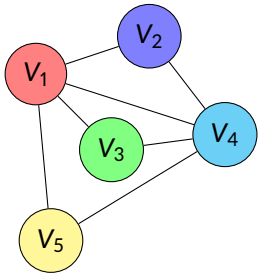
$$O(X, Y) = \frac{|X \cap Y|}{\min(|X|, |Y|)}.$$

- Evaluation based on real distribution of MPs in parties (≥ 0.9, several measures).

Party	ID
SYRIZA	0
DHSY	1
POTAMI	2
EK	3
ND	4
ANEL	5
XA	6

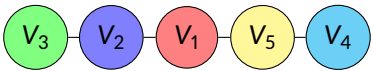


The Minimum Linear Arrangement Problem



Definition

Find an arrangement $\phi : V \to [1, 2, \ldots, n]$ such that its cost $c(\phi)$ is minimized.

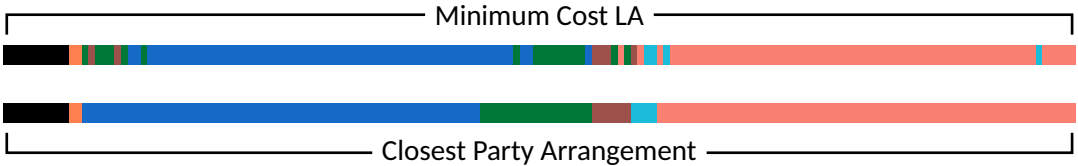


$$c(\phi) = \sum_{(u,v) \in E} w_{uv} \cdot |\phi(u) - \phi(v)|$$

Novel application of MinLA in SNA

The Minimum Linear Arrangement Problem Application

- Application: Local search MinLA algorithm in the NOI projection.
- Hypothesis: Members of the same party will appear consecutively in the MinLA of the NOI projection.



- Effectiveness: Over 90% of the theoretical max Kendall tau-b correlation coefficient.
- Finding: Closest party arrangement coincides approximately with right-to-left political spectrum.

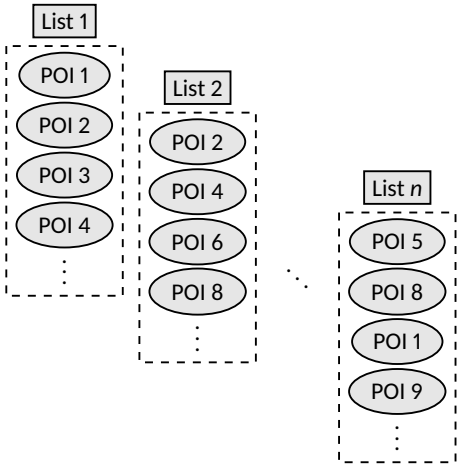
Part II: User Generated POI Lists

What is the criteria that social network users employ to group POIs in lists and how does it help recommender systems.

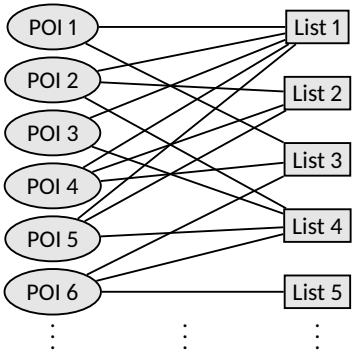
Overview

- Study the information encoded in user generated POI lists from LBSNs.
  - LBSN: Location Based Social Network.
- Utilize POI lists to drive a personalized recommender system.
  - List is a collection of related POIs.
  - Estimate the similarities among the POIs.
- Evaluate using online user survey and offline experiment.

Representation of POI Lists



POI list collections



Bipartite graph representation

Hypotheses & Intuition

Foursquare lists

- Foursquare LBSN: open and public access to lists through API.
- Future check-ins (To-Visit lists).

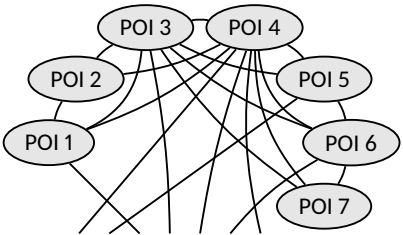
Advantages of the use of lists

- People often want to visit more places than they actually do.
- POIs are categorized based on specific criteria (location, time).
- Passive, no effort: users create lists for their own reasons.
- Adding a POI on a list is a conscious decision, location history may be inconsistent.
- No royalty or privacy issues.
- Purely structural: can be applied on fields where the lists are implicit.

Intuition

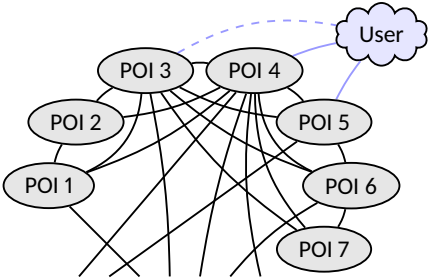
- Users intentionally group collections of POIs.
- POIs in a list are, by at least one relevant measure, related to one another.

Methodology



Bipartite projections

- Pairwise similarities among the POIs
- Link prediction // proximity
  - Set theoretic, graph theoretic and statistical similarity measures



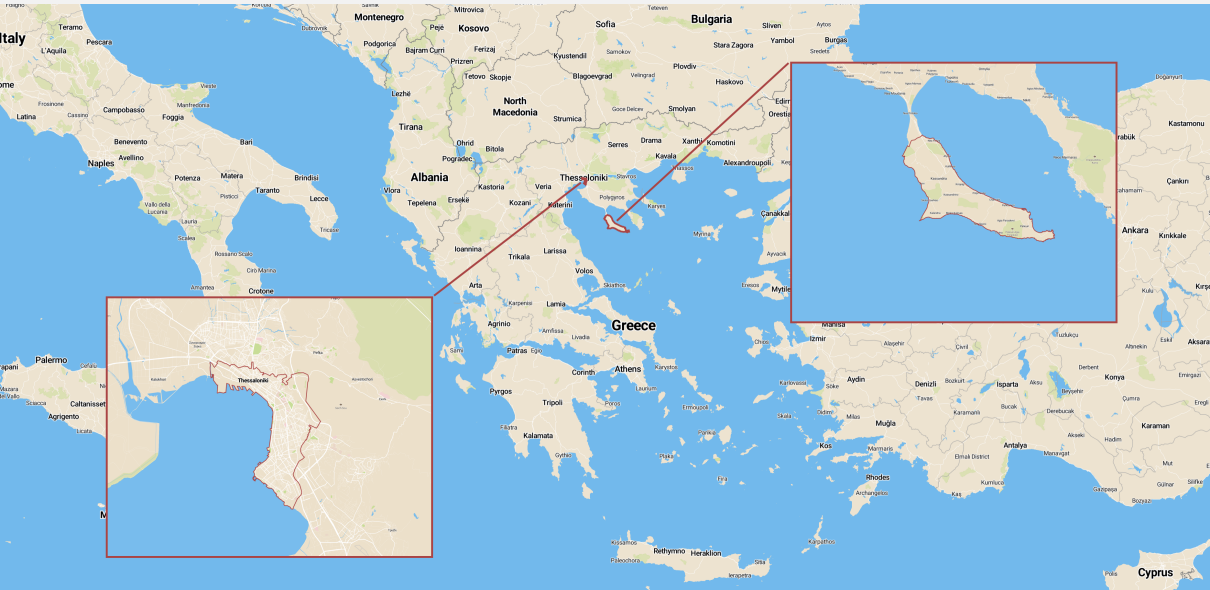
Recommendation method

- Recommendation score for any POI q: $w(q) = \sum_{i \in P} P(i)S(i,q)$
- Recommend the POIs that appear most similar to the POIs that the user finds interesting

POI Similarity Measures

β	Name	Description
aa	Adamic	Adamic/Adar index
is	Intersection	Number of common lists
jac	Jaccard	Intersection over union
ka	Modified MI	Modified mutual information
ku	Kulczynski-2	Intersection over harmonic mean
mi	MI	Mutual information of sets
cos	Ochiai	Cosine similarity or intersection over geometric mean
ov	Overlap	Intersection over minimum
ρ	Phi	Pearson correlation coefficient
sr	SimRank	Iterative calculation of SimRank
f1	Sørensen	Sørensen–Dice index or F1 score or intersection over arithmetic mean

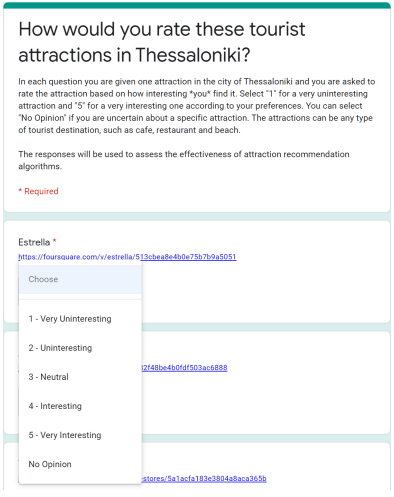
Case Study Areas

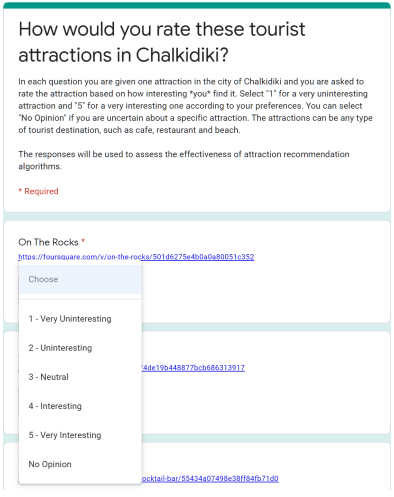


Experiments & Findings

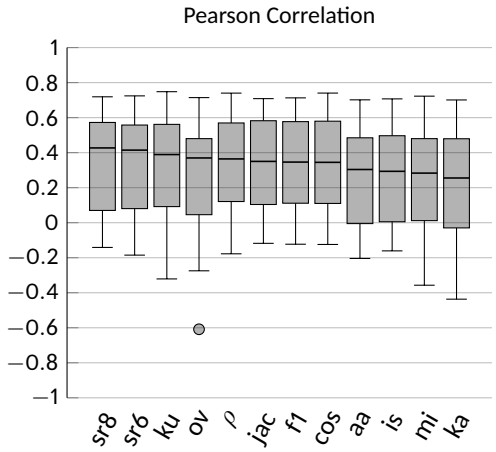
- Data retrieved from Foursquare (Foursquare lists).
  - Retrieved at mid 2020, lists span from 2011.
- Evaluation
  - Online evaluation based on user survey.
  - Offline evaluation by using a list as a profile.
- Significant results and above popularity (degree, likes, rating).
- Global similarities (SimRank, Adamic) might be more effective.

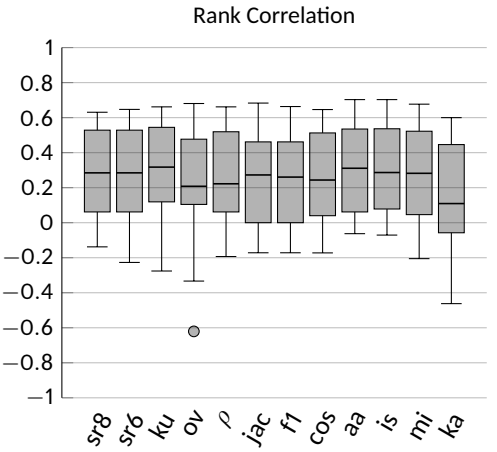
Online Evaluation: User Survey (19 + 11 POIs)



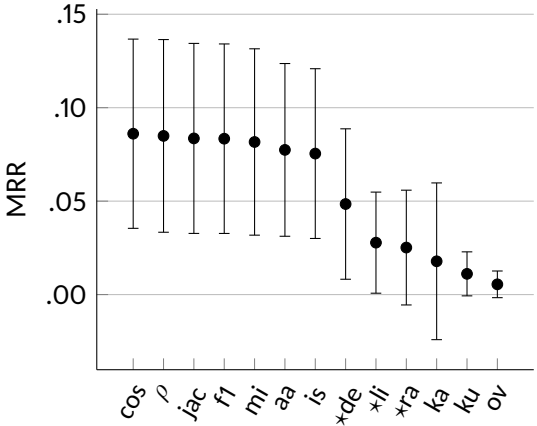


Online Evaluation (Thessaloniki – 28 Users)



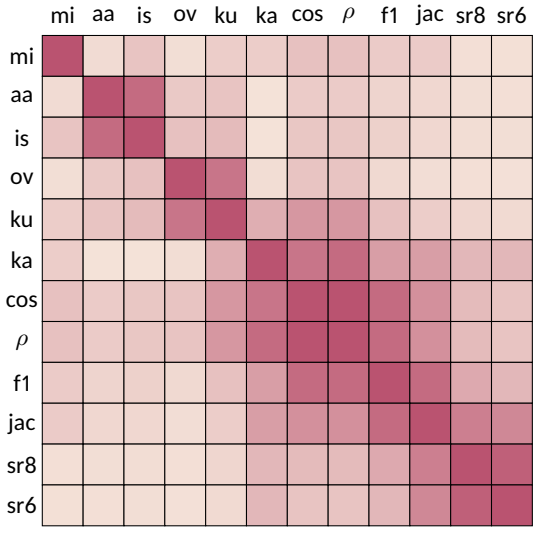


Offline Evaluation (747 Virtual Profiles)



- Baselines (⋆degree, ⋆likes, ⋆rating): profile ignored.
- Average of the average rank of missing POI for Ochiai (cos) was 17.75 (out of 3,397 POIs).
- There does not appear to be a similarity measure with clear advantages.

Correlation Among Projection Measures



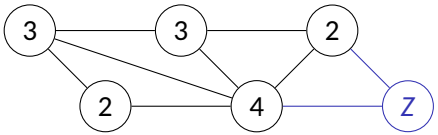
- Groups of measures with similar properties.
- Correlations with physical properties:
  - Geographic distance: mi, ka.
  - Categories: sr6, sr8, jac.
  - Rating difference: mi, ka.

Part III: Preferential Attachment

The mechanism under which users often select connections in social networks.

The Barabási–Albert model

- Barabási–Albert (BA) model: Preferential attachment (PA) that generates power law degree distributions on graphs.
- Power laws: Very often in nature, social networks, computer networks, software dependency graphs, financial networks, biological networks, airline networks.
- Conditions:
  1. Growing. New vertices enter the network and connect to m old vertices.
  2. Preferential. The probability of connecting to old nodes is proportional to their degrees.



$Z =$ newborn node

$m = 2$

The BA Model as a WRS application

- PA is an application of weighted random sampling (WRS).
  - But there are many WRS designs . . . Which one for $m > 1$?
- "Probability proportional to degree":
  - Which probability? *Inclusion*, *selection* or *independent*? Or something else?
- Indirectly defined as the inclusion probability.
  - First order inclusion probability (π_i) proportional to degree.
  - Probability of element i to exist in the sample after the selection process.
  - Higher order ($\pi_{ijk\ldots}$) probabilities not specified at all.

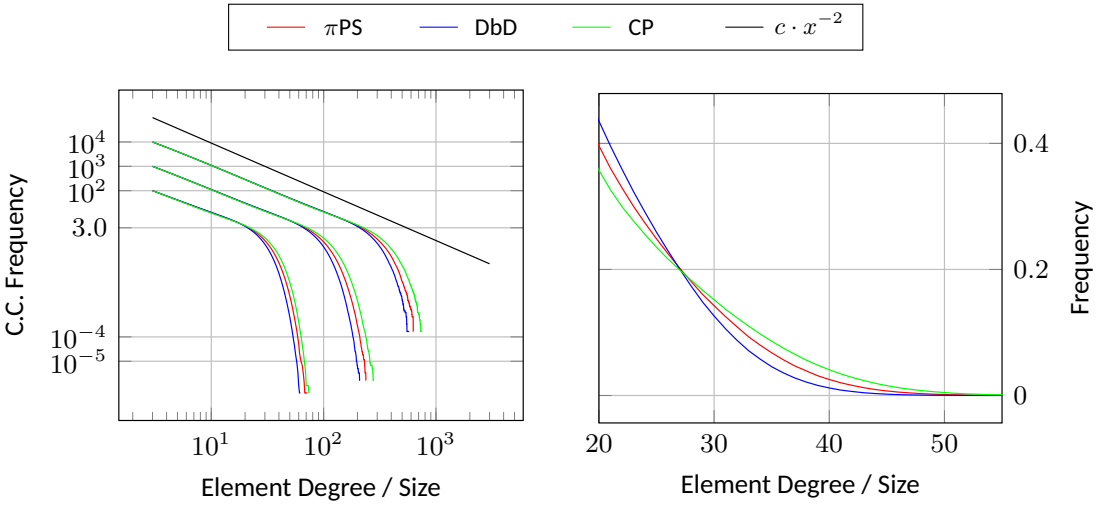
Current State: Approximate Model or Multigraph?

- Most existing models are either approximate or create multigraphs.
- Typically, the draw-by-draw procedure is utilized.
  - The inclusion probabilities are only approximately proportional to the degrees due to rejections.
  - Multigraph without rejections.
- Most notable software frameworks (NetworkX, iGraph, etc) implement both procedures.
- Higher order probabilities are usually ignored.

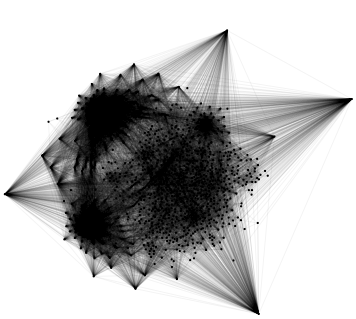
Draw-by-draw scheme

- Select one vertex with probability proportional to degree and keep it.
- Select another vertex with probability proportional to degree. If it's the same vertex ...
  1. Reject it or
  2. Create multigraph

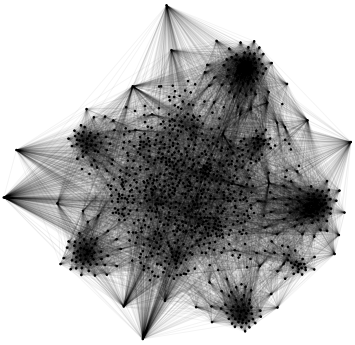
Degree Distribution (First Order)



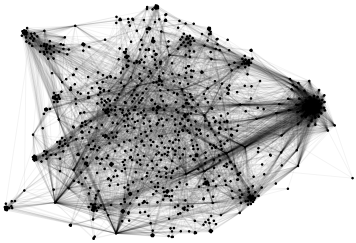
Jaccard Projection of Systematic Designs (Higher Order)



Random systematic



Ordered systematic



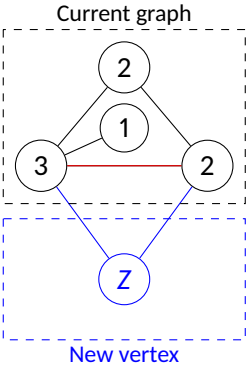
Romantic systematic

The SE Algorithms

- We designed a family of algorithms that interpret preferential attachment under the new perspective of random sampling.
  - All algorithms run in linear time with respect to the order of the graph n.
  - The inclusion probability of vertices is exactly proportional to their degree.
- Algorithms:
  - SE-A: Basic version that works for $m = 2$ and demonstrates the principle of operation.
  - Generalizations for $m > 2$: SE-B and SE-C, with different higher order properties.
- Degree distribution based on the original BA model (for large graphs):

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

The SE-A Algorithm

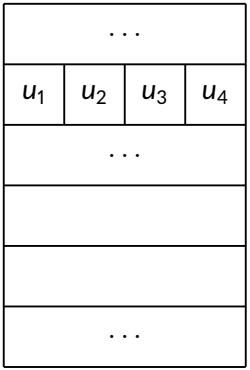


- Simple $\Theta(n)$ case for $m = 2$.
- Main loop (growth function):
  - Select one uniformly random existing edge.
  - Connect the new vertex with the ends of that edge.
- Each vertex exists in the edge set as many times as its degree.
  - Strict proportionality in the inclusion probabilities.
- Higher order: not all pairs can gain common neighbor.
  - 100% probability of forming a triangle after insertion.
  - $C(d) = 2/d \quad \bullet \quad C = 2\pi^2 - 19 \approx 0.73921.$

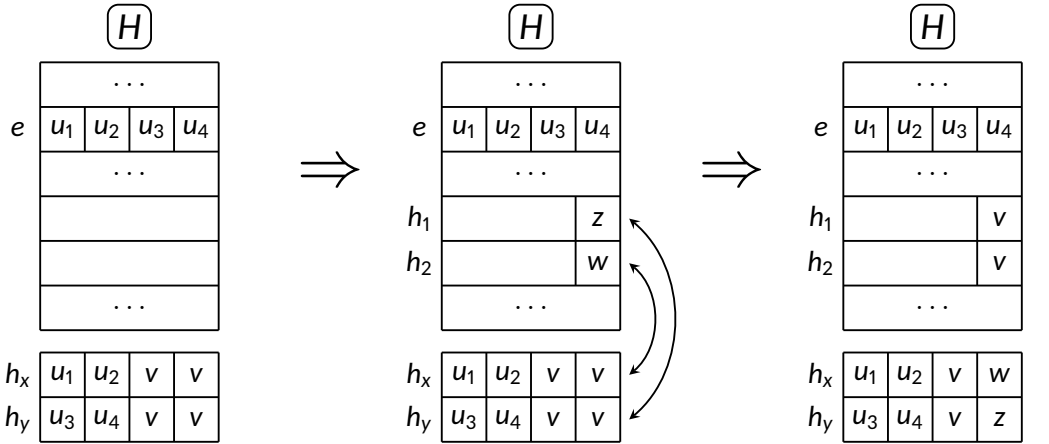
The SE-B Algorithm

- Generalization of SE-A for $m > 2$.
- Growth function:
  - Select one uniformly random row in H.
  - Connect the new vertex with all vertices in that row.
  - Update H such that the invariants are satisfied ...
- Auxiliary data structure H invariants:
  - No row can have duplicate vertices.
  - Each vertex must exist in as many rows as its degree.
- The H data structure resembles a (possibly non-simple) m-uniform hypergraph.
- The selection process resembles *whole sampling*.

H for $m = 4$



SE-B Algorithm Sketch (updating for $m = 4$)



Add all elements of e and m copies of v into the new hyperedges h_x and h_y.

Satisfy the invariants by swapping two copies of v with vertices in existing hyperedges.

Final state of the H list after the swap.

The SE-C Algorithm: Principle

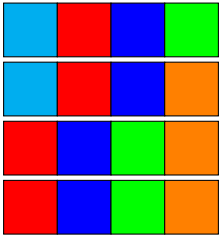
- Since most of the elements of (m) existing hyperedges are traversed (in the worst case), why not shuffle them?
  - Place the m^2 elements in a bag and shuffle them into m hyperedges such that the invariants are not violated.
  - Adjusting higher order inclusion probabilities at the same time.

Algorithm SE-C: Random Systematic Partitioning ($m = 4$)

Flatten 4 hyperedges of 4 elements each:

Shuffle the distinct elements and place consecutively:

Create 4 hyperedges via filling by column:



The SE-C Algorithm: Correctness & Performance

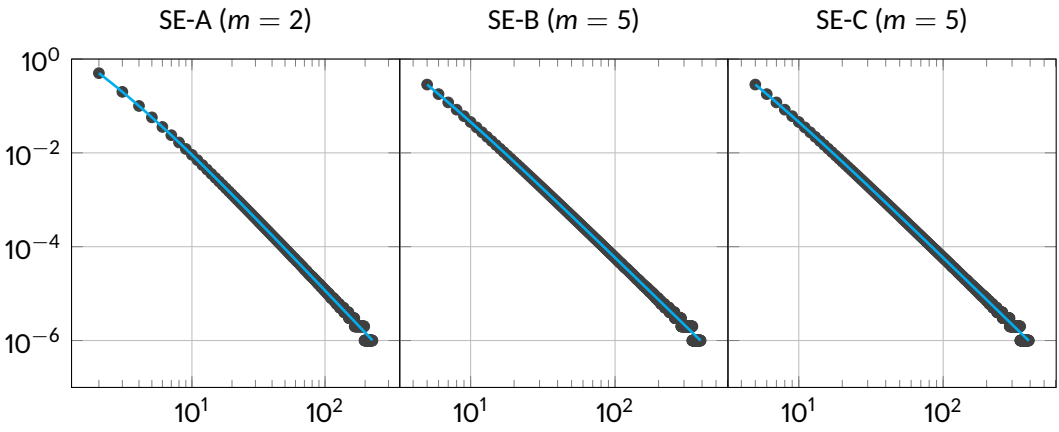
- Correctness: Only additional swaps are performed with respect to SE-B.
- Performance: Linear with respect to the order of the graph $\Theta(nm^2)$.
  - Each step is independent of n.

The SE-C Algorithm: WRS Relation

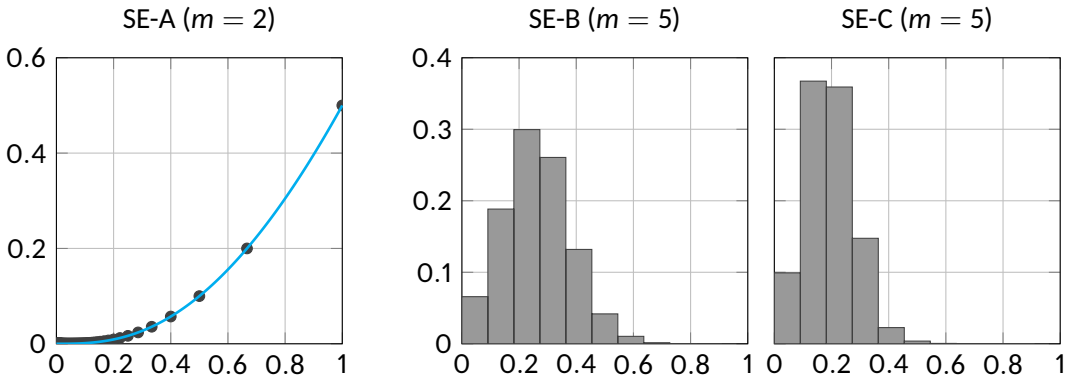
The role of random sampling in SE-C (and preferential attachment):

1. Select $m - 2$ existing hyperedges from the population of hyperedges in H.
2. Shuffle the node copies inside H using random systematic partitioning.
3. Whole sampling method to draw a sample (hyperedge) in constant time and update H to maintain proportionality of the inclusion probabilities.

SE Algorithm Degree Distribution



SE Algorithm Clustering Coefficient Distribution



Summary

Objective

Applications of the knowledge obtained through the behavior of OSN users via structural analysis.

Contributions to SNA

I How social network users can be used to infer the political orientation of other users.

II How social network users can accumulate knowledge through the POI lists to drive a personalized recommendation system.

III Fragmentation on the preferential attachment mechanism under which new users expose themselves to old users in social networks.

Future Work

- Structural analysis & user behavior knowledge is very promising $\rightarrow$ seek other perspectives.
- Utilize alternative methods of SNA (MinLA and others) $\rightarrow$ seek other methods.

Publications

- G. Stamatelatos, S. Gyftopoulos, G. Drosatos, P. S. Efraimidis. Revealing the political affinity of online entities through their Twitter followers. Information Processing & Management (2020). Elsevier.
  - G. Stamatelatos, S. Gyftopoulos, G. Drosatos, P. S. Efraimidis. Deriving the political affinity of Twitter users from their followers. SocialCom (2018).
- G. Stamatelatos, G. Drosatos, S. Gyftopoulos, H. Briola, P. S. Efraimidis. Point-of-interest lists and their potential in recommendation systems. Information Technology & Tourism (2021). Springer.
- Dataset, projections, other contributions
  - S. Gyftopoulos, G. Drosatos, G. Stamatelatos, P. S. Efraimidis. A Twitter-based approach of news media impartiality in multipartite political scenes. Social Network Analysis and Mining (2020). Springer.
  - H. Briola, G. Drosatos, G. Stamatelatos, S. Gyftopoulos, P. S. Efraimidis. Privacy leakages about political beliefs through analysis of Twitter followers. PCI (2018).
  - P. S. Efraimidis, G. Drosatos, A. Arampatzis. G. Stamatelatos, I. N. Athanasiadis. A Privacy-by-Design Contextual Suggestion System for Tourism. Journal of Sensor and Actuator Networks (2016). MDPI.
- G. Stamatelatos, P. S. Efraimidis. About Weighted Random Sampling in Preferential Attachment Models. `https://arxiv.org/abs/2102.08173`.
- G. Stamatelatos, P. S. Efraimidis. Whole Sampling Generation of Scale-Free Graphs. `https://arxiv.org/abs/2110.00287`.