Link Prediction with Combinatorial Structure: Block Models and Simplicial Complexes

Jon Kleinberg

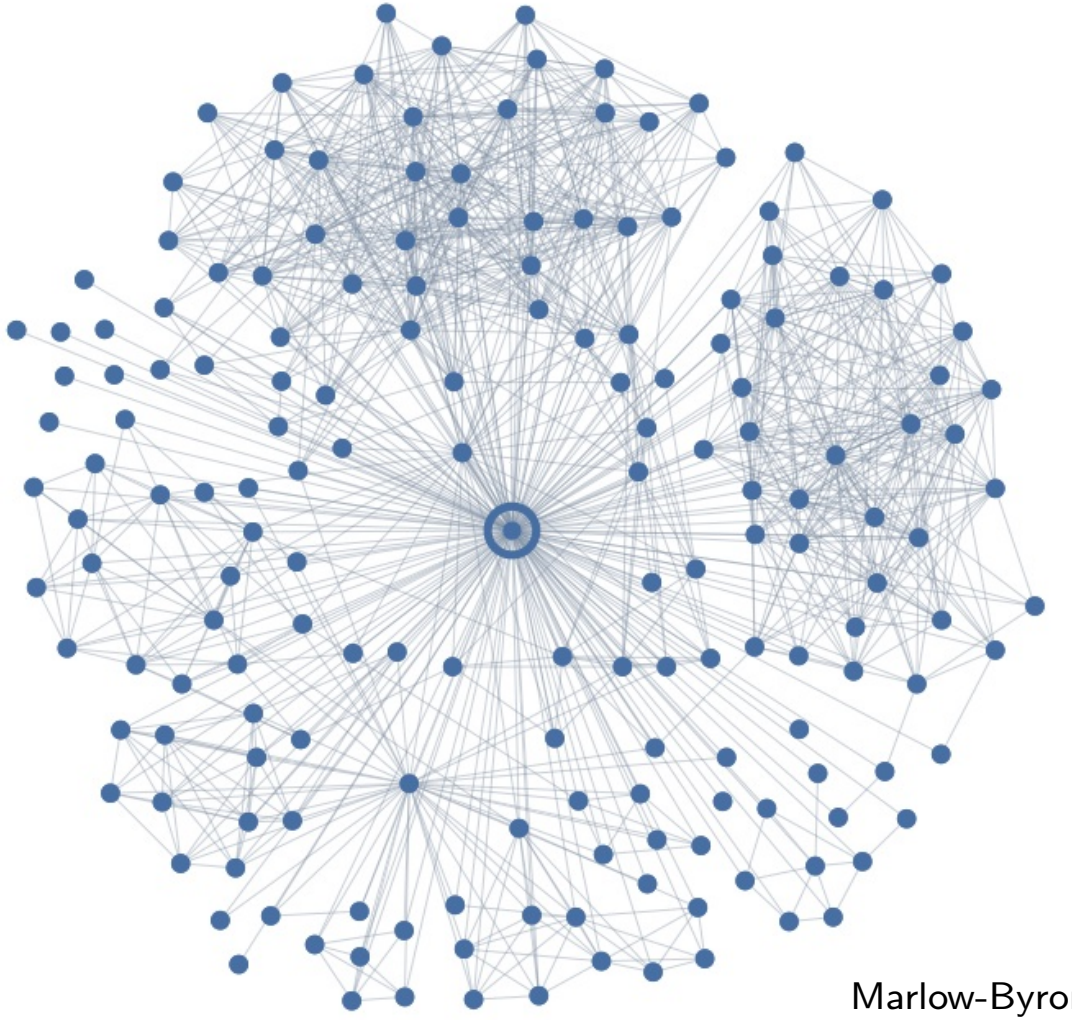
Including joint work with Rediet Abebe, Austin Benson, Ali Jadbabaie, Isabel Kloumann, Michael Schaub, and Johan Ugander.

Cornell University



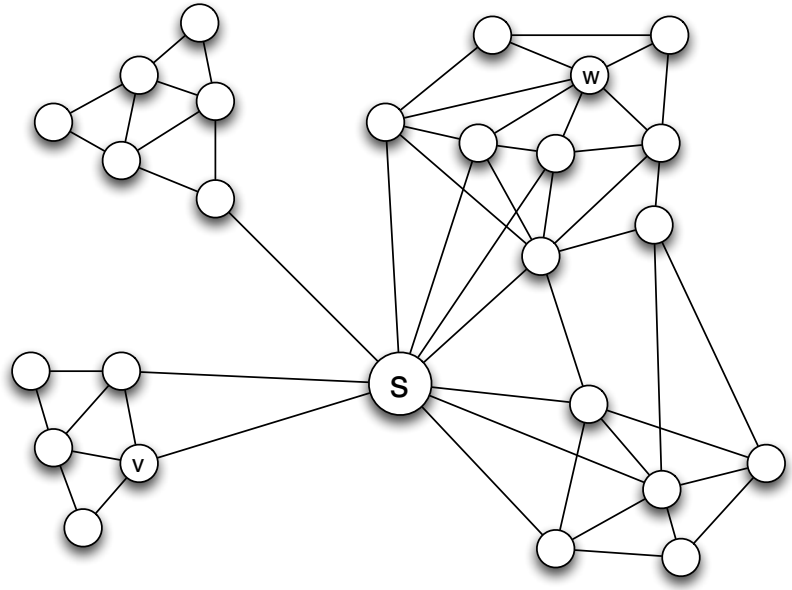


Neighborhoods and Information-Sharing



Marlow-Byron-Lento-Rosenn 2009

Node neighborhoods are the "input" in the basic metaphor for information sharing in social media.



A definitional problem: For a given m, who are the m closest people to you in a social network?

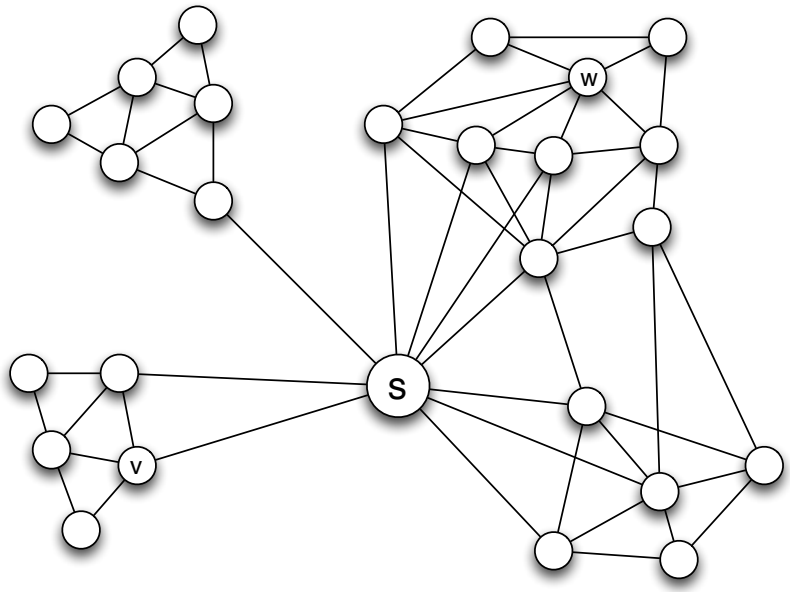
Defining "closeness" just by hop count has two problems:

- Not all nodes at a given number of hops seems equally "close."
- The small-world phenomenon: most nodes are 1, 2, 3, or 4 hops away.

Many applications for this question:

- Content recommendation: What are the m closest people discussing?
- Link recommendation: Who is the closest person you're not connected to?
- Similarity: Which nodes should I include in a group with s?

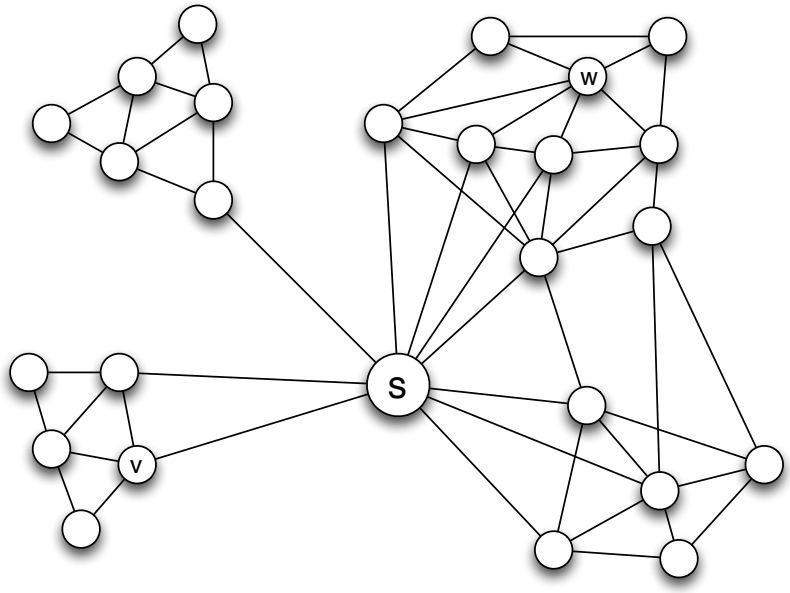
Problems Related to Proximity



Two testbed versions of these questions:

- Link Prediction: Given a snapshot of the network up to some time t_0, produce a ranked list of pairs likely to form links after time t_0. [Kleinberg-LibenNowell 2003, Backstrom-Leskovec 2011, Dong-Tang-Wu-Tian-Chawla-Rao-Cao 2012].
- Seed Set Expansion: Given some representative members of a group, produce a ranked list of nodes likely to also belong to the group [Anderson-Chung-Lang 2006, Abrahao et al 2012, Yang-Leskovec 2012].

Problems Related to Proximity

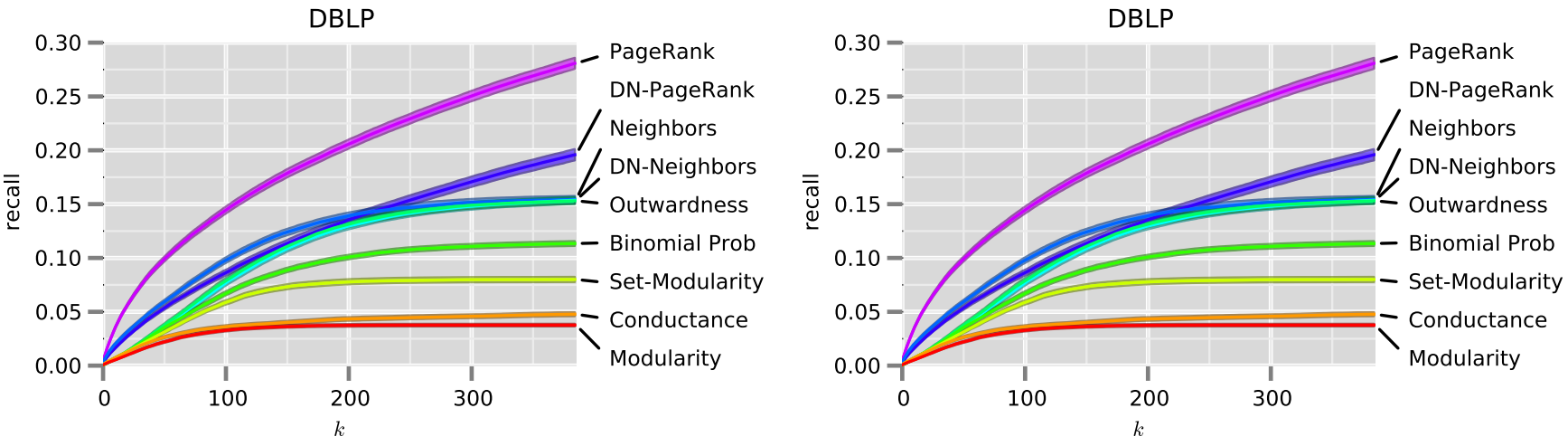


Methods for seed set expansion

- Local: Add nodes iteratively by some measure of the number of edges they have to current group members. [Clauset 2005, Bagrow 2008, Mehler-Skiena 2009, Mislove et al 2010].
- Non-local: Add nodes iteratively using a measure based on more than just direct connections.

.

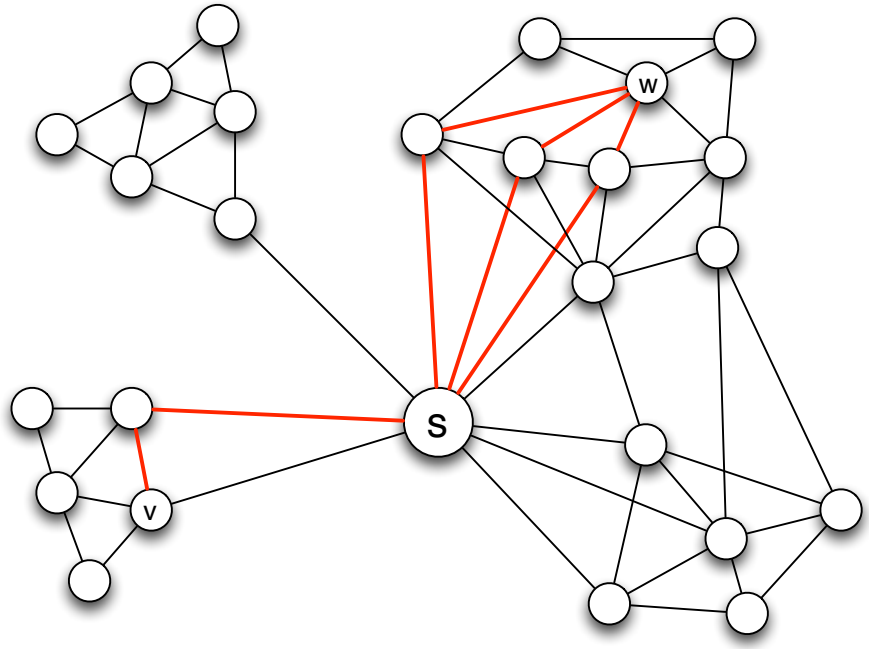
Some Results on Seed Set Expansion



Kloumann-Kleinberg 2014

Basic observations

- Personalized PageRank outperforms local methods:
  Random walk with probability $p > 0$ of resetting to start each step;
  rank nodes by stationary probability [Haveliwala 02, Jeh-Widom 03].
- Heat kernel methods (also non-local) achieve comparable performance
  [Kloster-Gleich 2014]



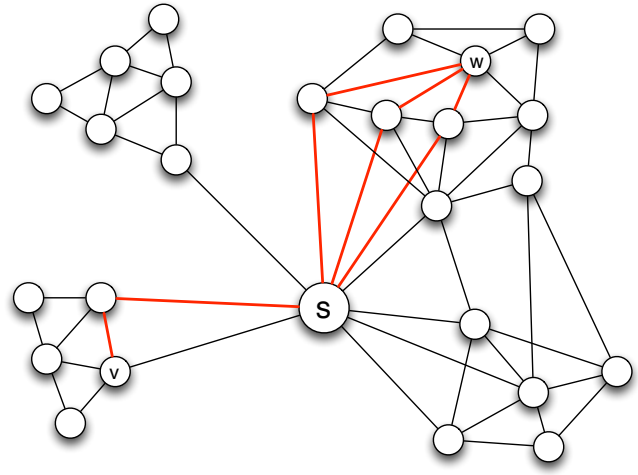
We want to find nodes "close" to a given s.

- A generic language for all methods discussed thus far:
  for each other node v, and $k = 1, 2, 3, ...$,
  count the number of k-step walks from s to v.
- For v in the figure, the counts are $(1, 1, 5, ...)$.
  For w in the figure, the counts are $(0, 3, 10, ...)$.

For node v and walk length k, let $r_k(v)$ be the probability that a random k-step walk from s ends at v.

- Map each node v to a vector $(r_1(v), r_2(v), r_3(v), ...)$.

Given a node v with vector $(r_1(v), r_2(v), r_3(v), ...)$, want to define a single number specifying closeness.



- Personalized PageRank of v [Haveliwala 02, Jeh-Widom 03]:

$$\alpha r_1(v) + \alpha^2 r_2(v) + \alpha^3 r_3(v) + \cdots$$

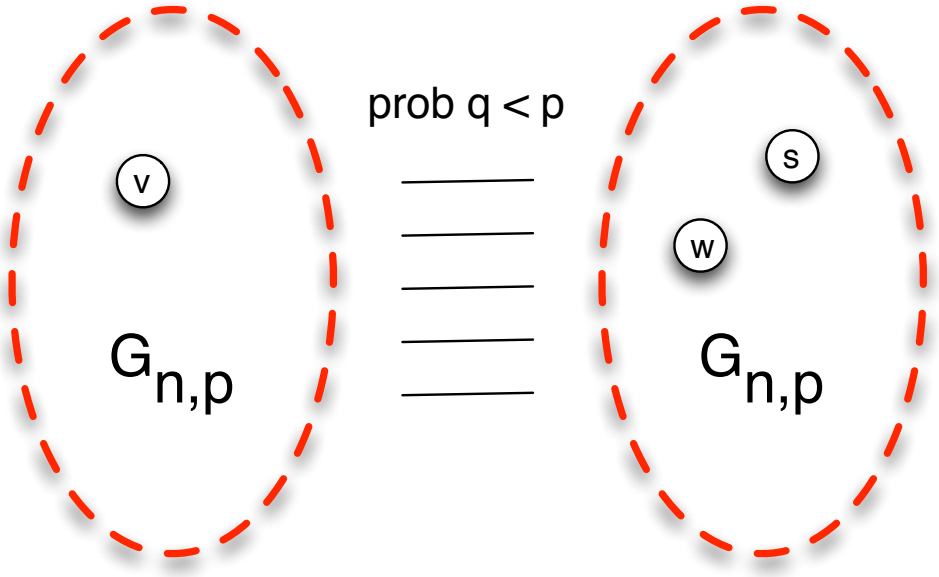
- Heat kernel value of v [Chung 07, Kloster-Gleich 14]:

$$\left(\frac{t^1}{1!}\right) r_1(v) + \left(\frac{t^2}{2!}\right) r_2(v) + \left(\frac{t^3}{3!}\right) r_3(v) + \cdots$$

These are just rules of the form $w_1 r_1(v) + w_2 r_2(v) + w_3 r_3(v) + \cdots$ for a choice of weights $w_1, w_2, w_3, ...$

- Is there a principled way of deciding which are the "right" weights to use?

Stochastic Block Models



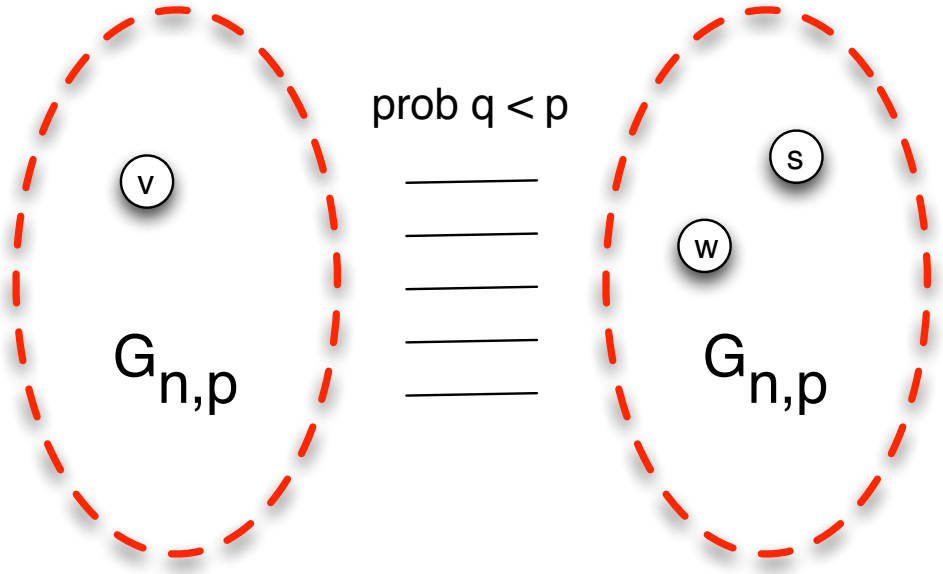
Stochastic block model (SBM)

[Holland et al 83, Dyer-Frieze 89, Condon-Karp 01, McSherry 01]

- Traditional problem statement: recover planted partition.
- Related problem statement: recover partition positively correlated with truth [Decelle et al 11, Mossel et al 12, Massoulié 13, Abbe et al 14]

But we can also use the SBM to evaluate different ways of weighting walks.

Stochastic Block Models



Stochastic block model (SBM)

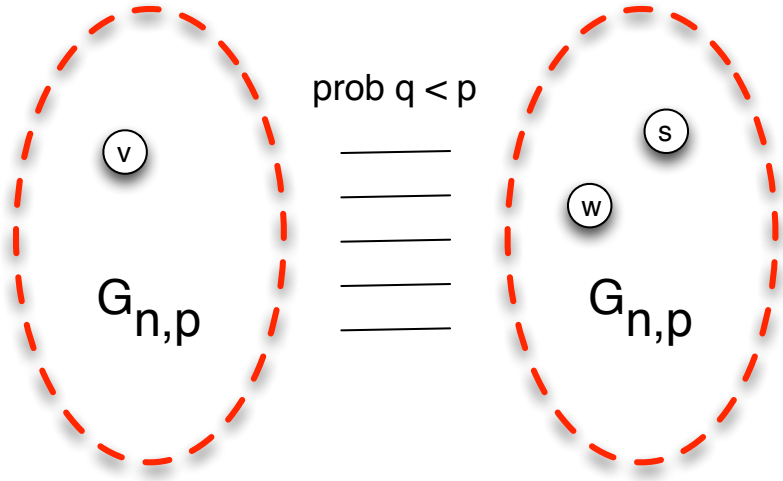
[Holland et al 83, Dyer-Frieze 89, Condon-Karp 01, McSherry 01]

- Generates random input graph with a "hidden" correct answer.
- For each node v, we get a vector $(r_1(v), r_2(v), r_3(v), ..., r_\ell(v))$.

Leads to a natural classification problem: what is the optimal linear separator for the vectors in the two blocks?

[Kloumann-Ugander-Kleinberg 2017]

- Generate vectors $(r_1(v), r_2(v), r_3(v), ..., r_\ell(v))$.
- Find optimal linear separator.



Theorem [Kloumann-Ugander-Kleinberg 17]:

For SBM with constant parameters $q < p$, and any $\varepsilon > 0$, there is a sufficiently large number of nodes n such that the optimal linear separator sorts the nodes by $\sum_{i=1}^{\ell} w_i r_i(v)$, where with high prob. $(w_1, w_2, \ldots, w_\ell)$ is ε-close to the personalized PageRank vector $(\alpha, \alpha^2, ..., \alpha^\ell)$.

(In particular, $\alpha = \left(\frac{p-q}{p+q}\right)$.)

Sketch of Proof

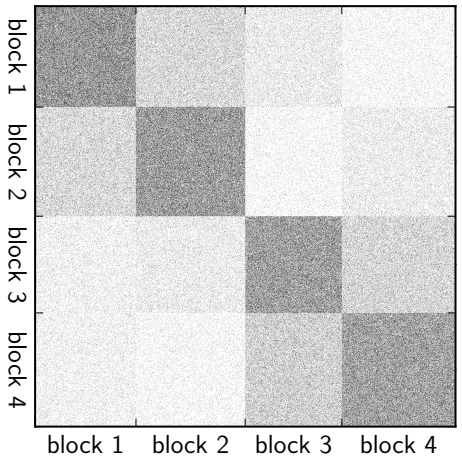
Theorem [Kloumann-Ugander-Kleinberg 17]:

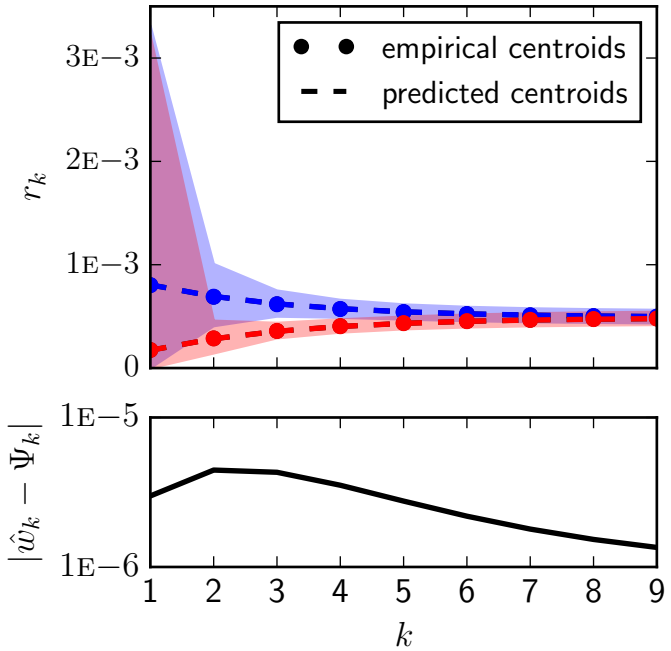
For SBM with constant parameters $q < p$, and any $\varepsilon > 0$, there is a sufficiently large number of nodes n such that the optimal linear separator sorts the nodes by $\sum_{i=1}^{\ell} w_i r_i(v)$, where with high prob. $(w_1, w_2, \ldots, w_\ell)$ is ε-close to the personalized PageRank vector $(\alpha, \alpha^2, ..., \alpha^\ell)$.

(In particular, $\alpha = \left(\frac{p-q}{p+q}\right)$.)

- Establish concentration of walk landing probabilities in each of the two blocks.
- Recurrence for the landing probabilities.
- Solution of recurrence yields a vector between means in the two blocks.

Extensions



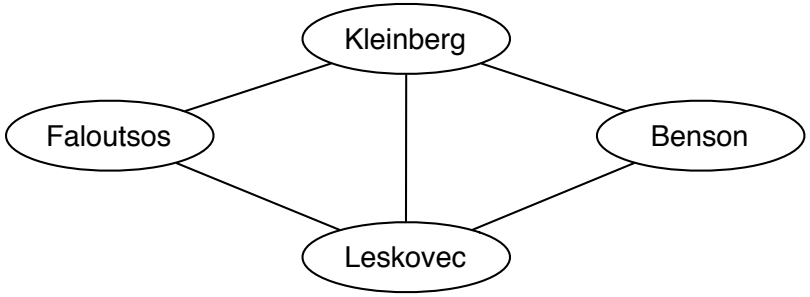


- Can compute optimal linear separator from a matrix recurrence for any SBM with $C > 2$ blocks partitioned into an in-class S and out-class T.
- With an equal-expected-degree condition (as in the case $C = 2$), personalized PageRank is still the optimal linear separator.

Further extension:

- Rather than a linear separator, improved performance using higher moments of vectors of landing probabilities.

Higher-Order Structure



In many settings where we use graphs to model interactions, we actually have collections of sets

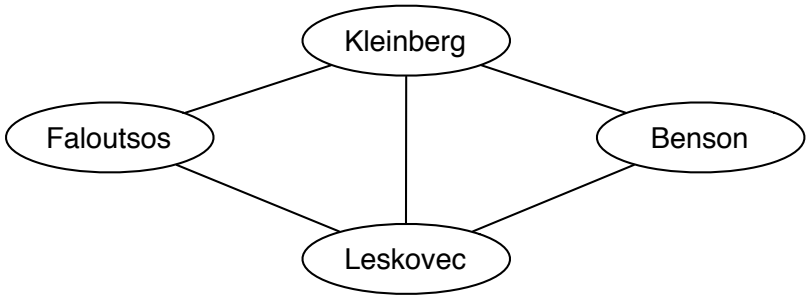
[Benson-Gleich-Leskovec 2016, Newman-Watts-Strogatz 2002].

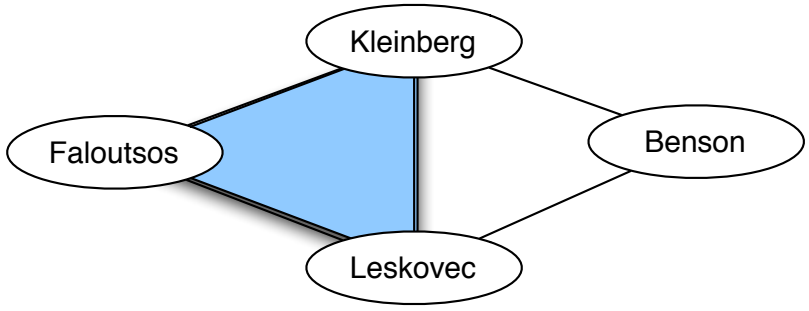
- Co-authorships occur in sets of more than two.
- Communication (e.g. e-mail) often goes to groups.
- Semantically related concepts occur in groups.
- Metabolic interactions often occur in sets of more than two.

Formalisms include hypergraphs, set systems, simplicial complexes, affiliation networks, but much is still unexplored.

- What can we use as a model problem with a clear objective function?

Higher-Order Structure





In many settings where we use graphs to model interactions, we actually have collections of sets

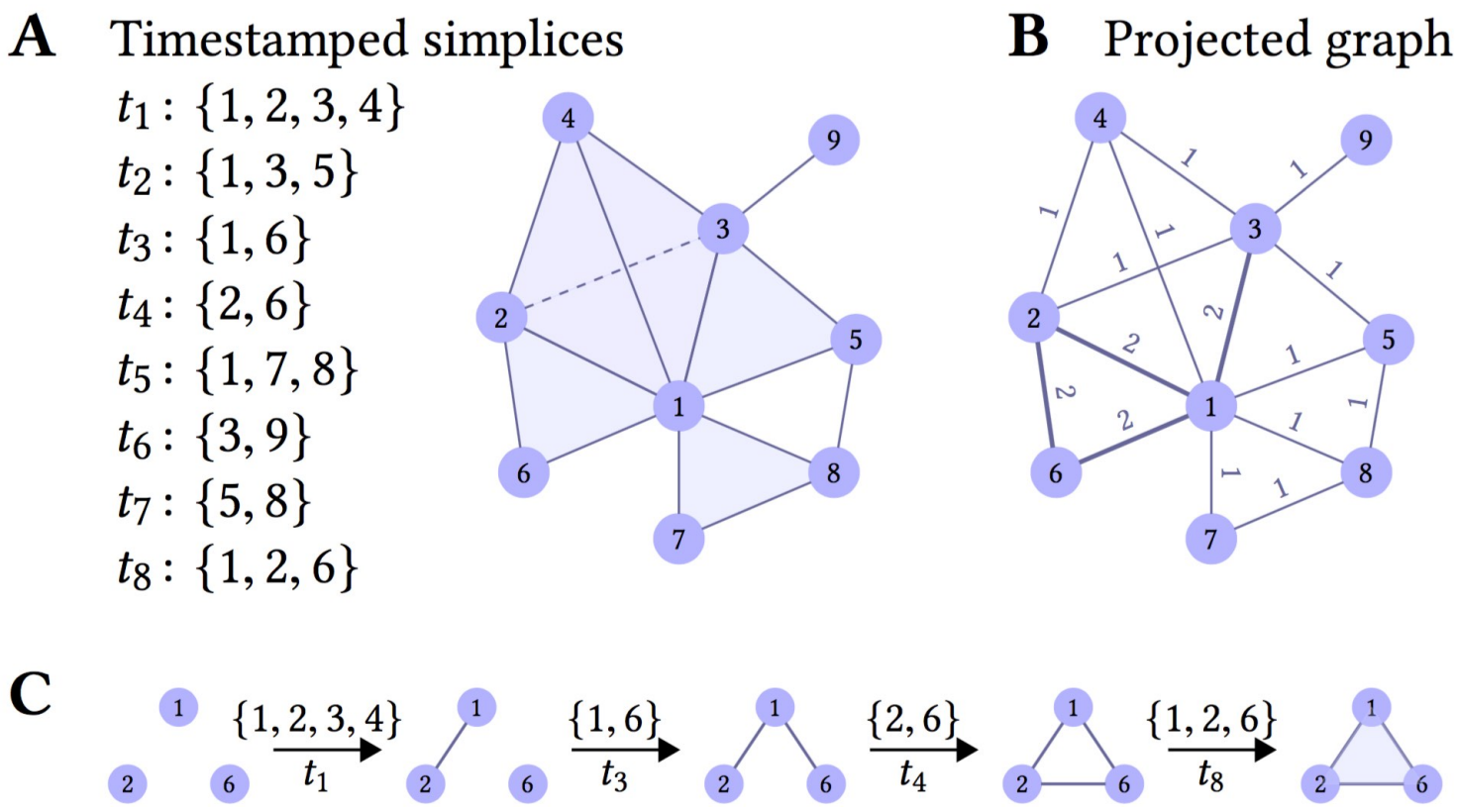
[Benson-Gleich-Leskovec 2016, Newman-Watts-Strogatz 2002].

- Co-authorships occur in sets of more than two.
- Communication (e.g. e-mail) often goes to groups.
- Semantically related concepts occur in groups.
- Metabolic interactions often occur in sets of more than two.

Formalisms include hypergraphs, set systems, simplicial complexes, affiliation networks, but much is still unexplored.

- What can we use as a model problem with a clear objective function?

Link Prediction

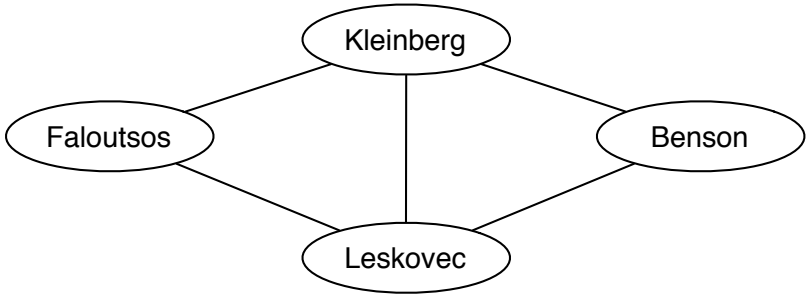


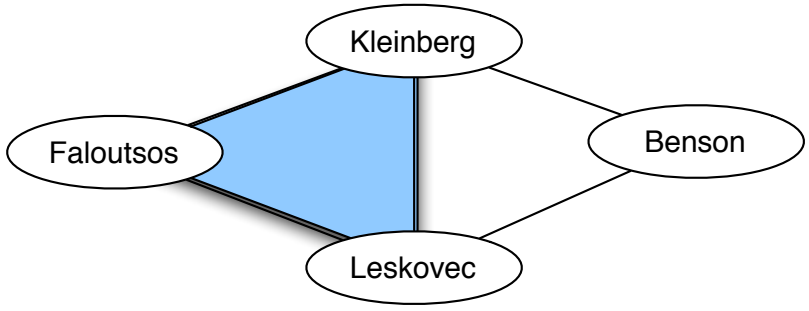
Higher-order link prediction problem

[Benson-Abebe-Schaub-Jadbabaie-Kleinberg 2018, https://github.com/arbenson/ScHoLP-Tutorial]

- Given a time-stamped sequence of sets up to time t_0, predict which sets will form in the future.
- Begin by focusing on sets of size three (triangles).

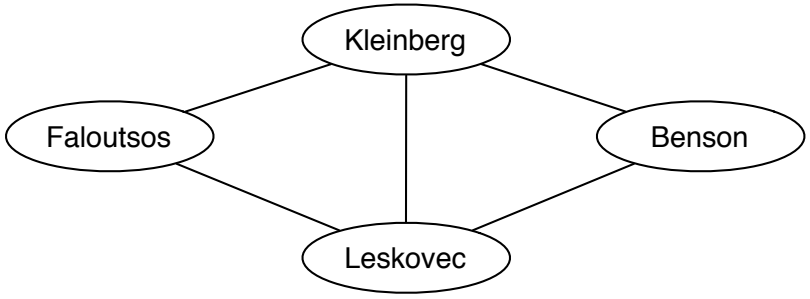
Simplicial Closure

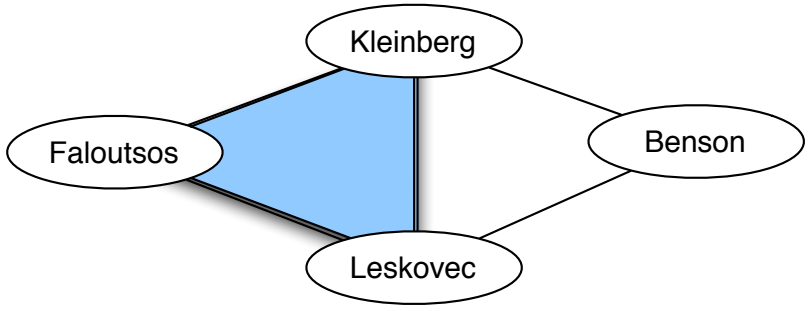




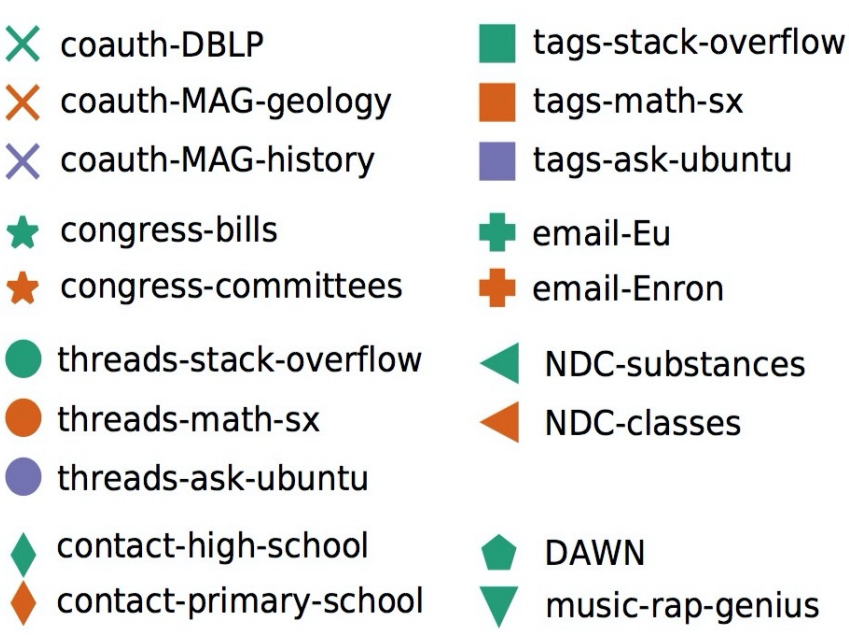
First Q: what are relative proportions of open and closed triangles?

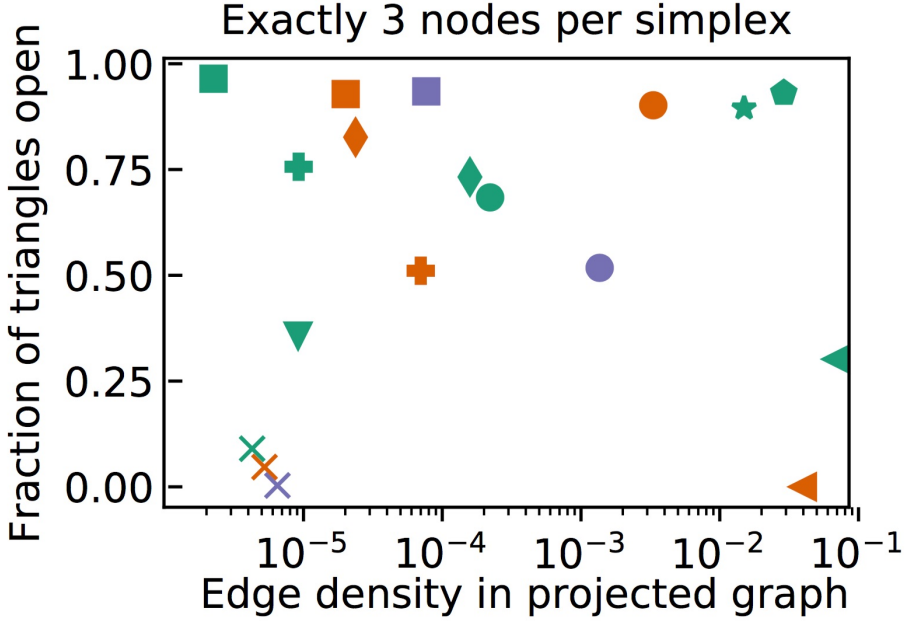
Simplicial Closure





First Q: what are relative proportions of open and closed triangles?

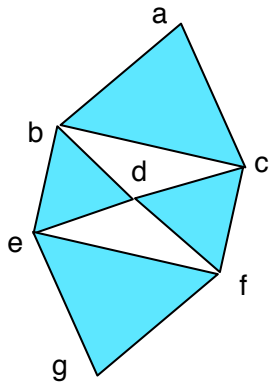




A Random Baseline

What's a reasonable random baseline for simplicial closure?
Insert triangle on each triple of nodes indep. with prob. n^{-b}.

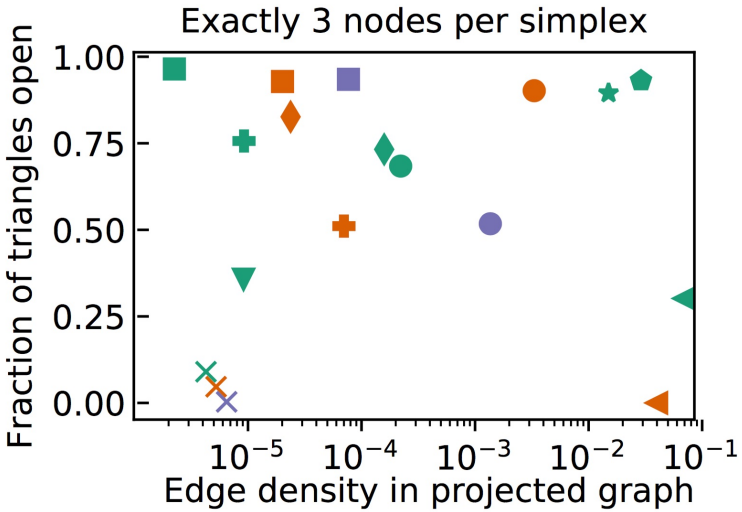
- Expected number of closed triangles is $\Theta(n^{3-b})$.
- For $b < 1$, almost all edges form, and so almost all triples without closed triangles have open triangles.
- For $b > 1$, number of open triangles is $\Theta(n^{3(2-b)})$.
- For $b > 3/2$, number of closed triangles grows faster.

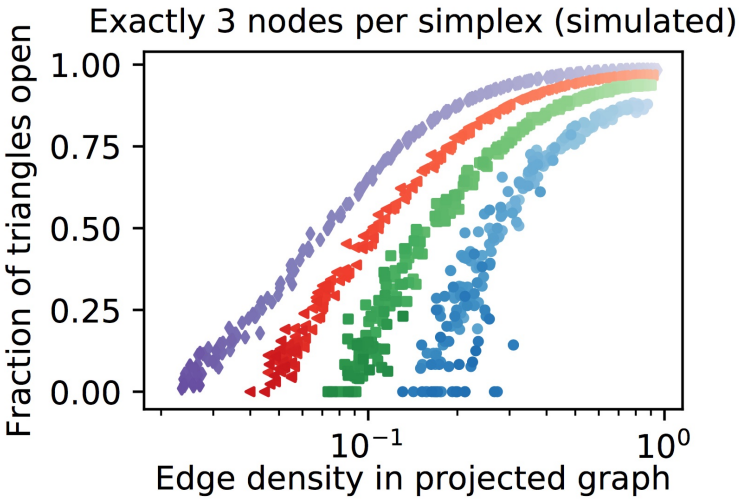


A Random Baseline

What's a reasonable random baseline for simplicial closure?
Insert triangle on each triple of nodes indep. with prob. n^{-b}.

- Expected number of closed triangles is $\Theta(n^{3-b})$.
- For $b < 1$, almost all edges form, and so almost all triples without closed triangles have open triangles.
- For $b > 1$, number of open triangles is $\Theta(n^{3(2-b)})$.
- For $b > 3/2$, number of closed triangles grows faster.

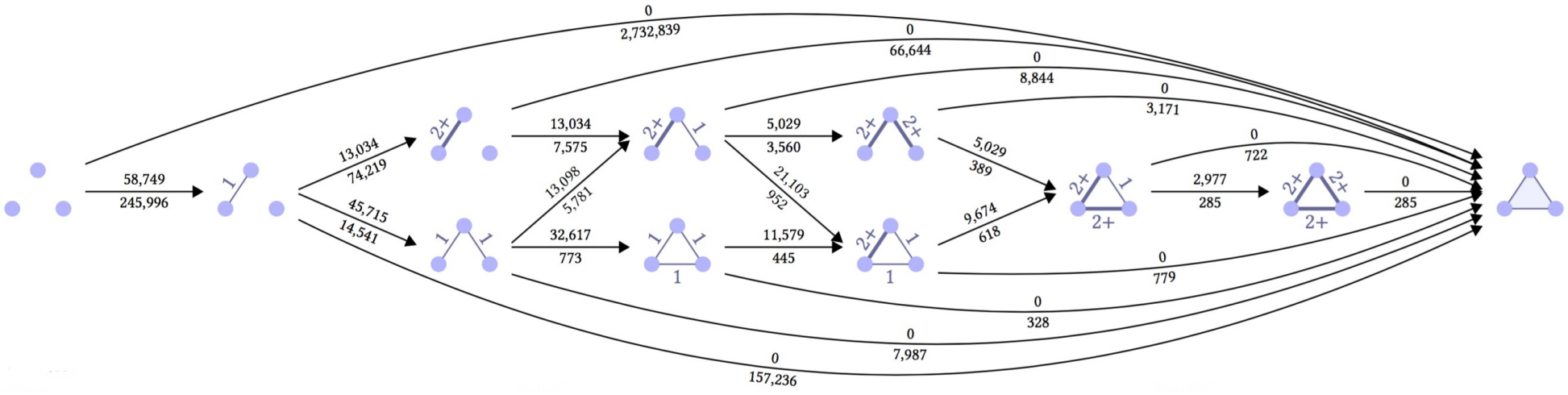




Temporal Dynamics

Can enumerate all possible trajectories to reach a closed simplex.

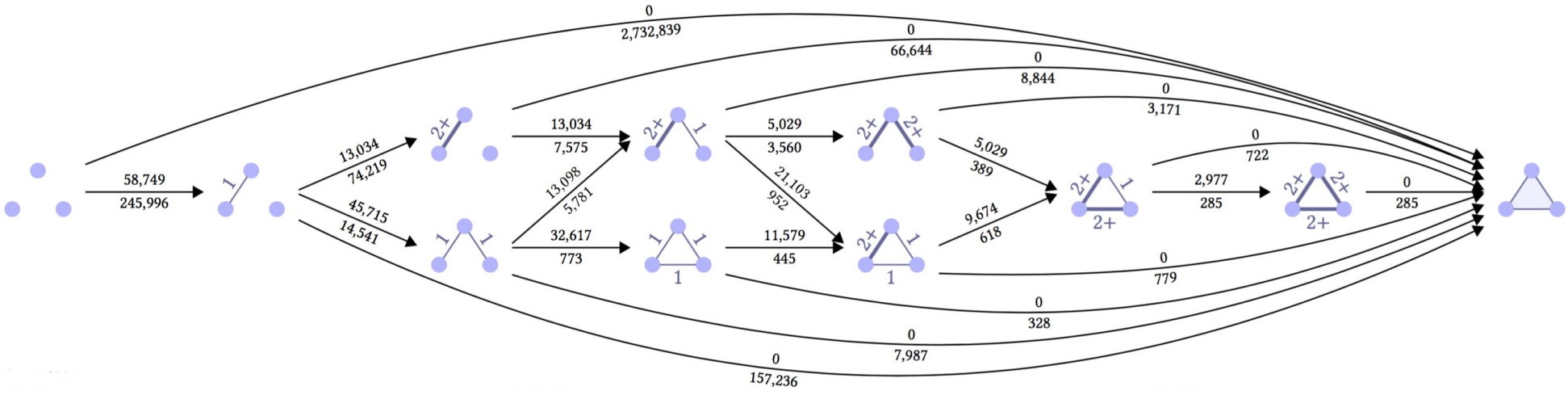
For example, in co-authorships on history papers:



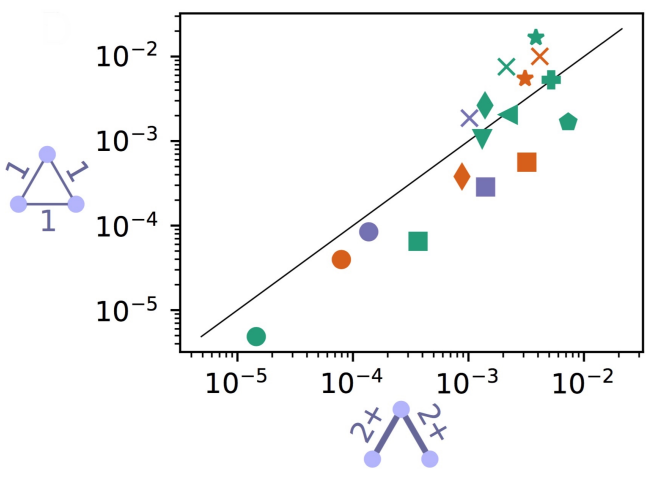
Temporal Dynamics

Can enumerate all possible trajectories to reach a closed simplex.

For example, in co-authorships on history papers:

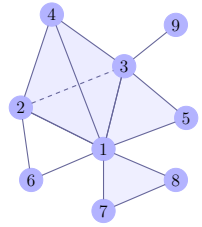


Compare closure probabilities for "strong wedge" and "weak open triangle."



Prediction Algorithms

Predict whether triangle $\{i, j, k\}$ will form.
Four categories of algorithms, based on different types of network information (plus combinations using supervised learning).



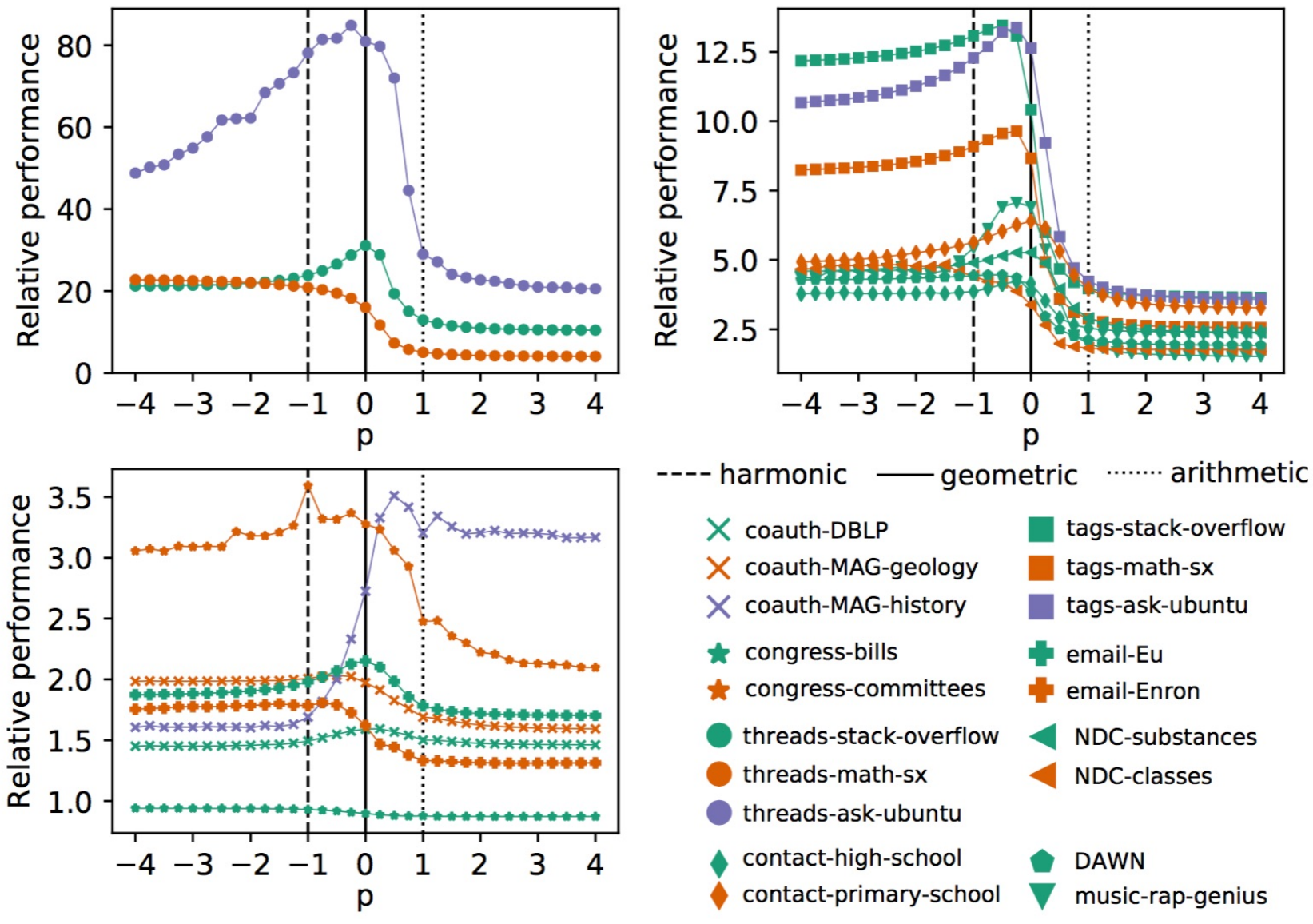
- Generalized means of edge weights $[(W_{ij}^p + W_{jk}^p + W_{ik}^p)/3]^{1/p}$
  (Note $p = -1$ is harmonic mean, and limit as $p \to 0$ is geometric mean.)
- Edge weights to common neighbors in projected graph.
- PageRank-like measures on projected graph.
- Generalizations of PageRank to simplicial complexes
  [Steenbergen-Klivans-Mukherjee 2014, Parzanchevski-Rosenthal 2016, Horn-Jadbabaie-Lippner 2017]

First category using just edge weights on $\{i, j, k\}$ comparable to all others (though supervised learning still improves).

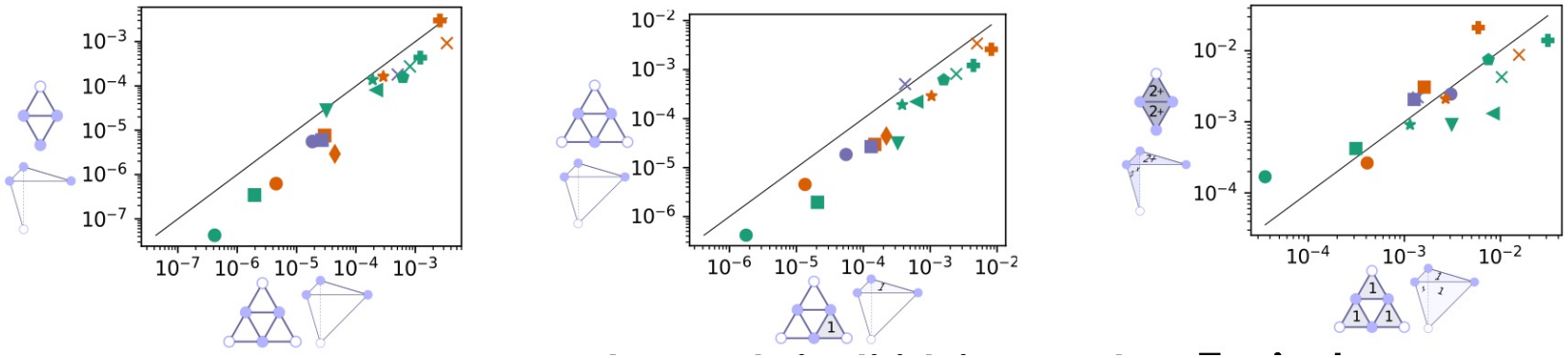
- No analogue for traditional pairwise link prediction.

Prediction Using Local Edge Weights

$$[(W_{ij}^p + W_{jk}^p + W_{ik}^p)/3]^{1/p}$$



Further Directions



- Can extend simplicial link prediction beyond sets of size three; with sets of size four, analogous contrast between lighter denser structures and heavier sparser structures.
  (Several new computational insights needed to scale this.)
- Larger sets raise fundamental scalability challenges.
- Can we find stronger non-local methods for prediction on sets of size > 2?
- Generative models with planted structure for sets of size > 2 as analytical testbed for higher-order link prediction?