Vertex Nomination Via Local Neighborhood Matching

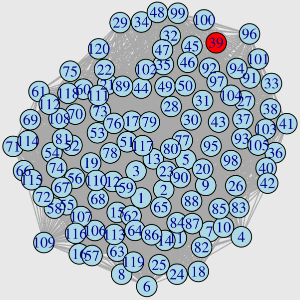
Heather G. Patsolic
In collaboration with: C.E. Priebe, V. Lyzinski, and Y. Park

Johns Hopkins University

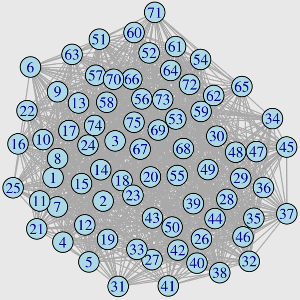
August 3, 2016

Problem Formulation

- Have two large networks on overlapping, non-identical vertex sets.
- There is a vertex of interest (VOI) in one network we'd like to identify in the other.



(a) Network A



(b) Network B

Challenge

- Often vertex attributes alone are not enough to identify VOI in the other network.
- Networks can be too large for graph matching to be efficient.

Challenge

- Often vertex attributes alone are not enough to identify VOI in the other network.
- Networks can be too large for graph matching to be efficient.

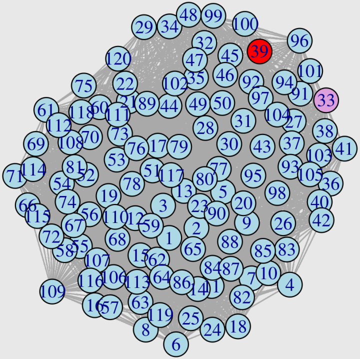
So how do we proceed?

Mathematical Framework for Simulations: ρ-SBM

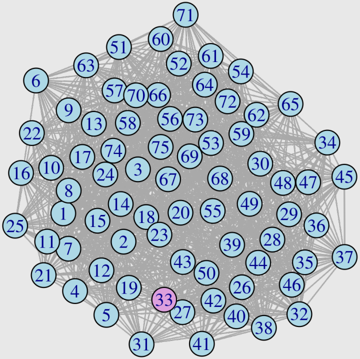
$G_1, G_2 \sim \rho-$SBM:

- Nodes are divided into groups.
- Probability of an edge existing between any pair of vertices in a graph depends only on the block membership of those vertices.
- Edges are marginally conditionally independent.
- Edge presence between vertices i and j in G_1 and vertices i and j in G_2 has correlation ρ.
- Otherwise, edge presence is independent across graphs.

Step 1: Acquiring Seeds



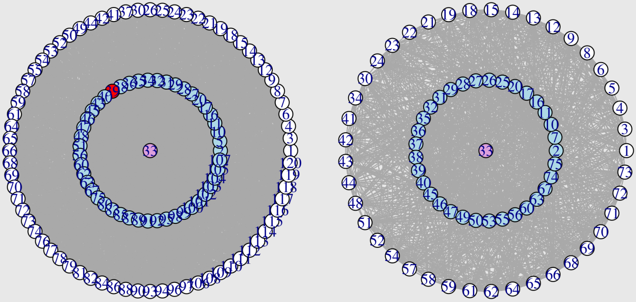
(a) Network A



(b) Network B

Figure: Find a vertex adjacent to VOI with verifiable corresponding vertex in second network (this is the initial *seed*).

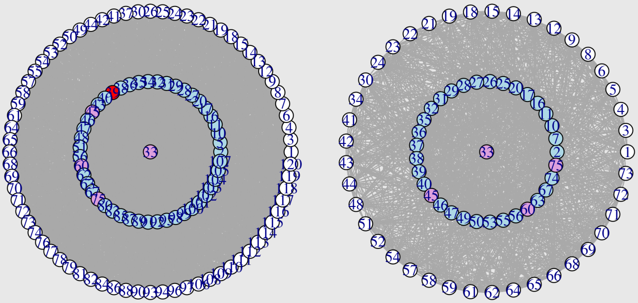
Step 1: Acquiring Seeds



(a) Network A (b) Network B

Figure: Generate h-hop neighborhood around this seed in both graphs. In this example, $h = 1$.

Step 1: Acquiring Seeds

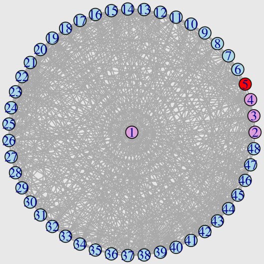


(a) Network A (b) Network B

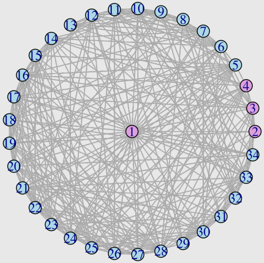
Figure: Find more seeds across these two induced subgraphs. Call the full seed sets S and S'.

Step 2: Finding Candidates

- $C'_x = \{5, \ldots, 34\}$ is the set of candidate (non-seed) vertices in the second induced subgraph.
- Note: if match to VOI is not in C'_x, we are doomed to failure. So be it; we still proceed. – Assume x' exists.

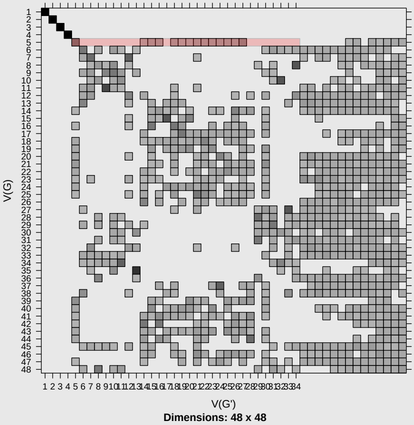


(a) *h*-hop neighborhood induced sub-network of A



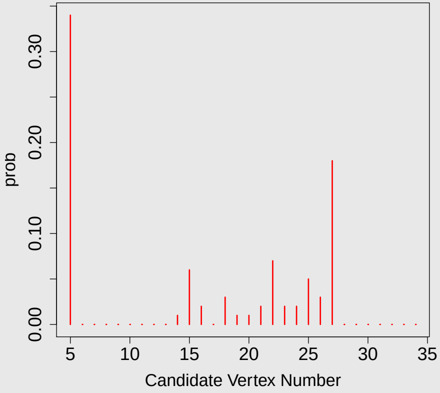
(b) *h*-hop neighborhood induced sub-network of B

Step 3: Matching Graphs



- Repeatedly use seeded graph matching (SGM), modified from [FAP12], to align the networks generated by these neighborhoods.
- Output a probability matrix P such that $P[i,j]$ is the proportion of times vertex j in second network mapped to vertex i in first network.

Step 4: Nominations for VOI



- The most likely nominate for the VOI is in $\arg\max_{v \in C'_x} P[x, v]$.
- Nomination list for the VOI, x, is the list of vertices in C'_x ordered from most to least probable.

VN via LNM algorithm

Input: Graphs G_1 and G_2; $x \in V(G_1)$; R; h

VN via LNM algorithm

Input: Graphs G_1 and G_2; $x \in V(G_1)$; R; h
Step 1: Find pair of initial seeds $s_1 \in V(G_1)$ and $s_1' \in V(G_2)$ so that s_1 is adjacent to x in G_1 .
Generate h-hop neighborhoods around initial seeds (be sure VOI is in first neighborhood).
Find more seeds if possible.

VN via LNM algorithm

Input: Graphs G_1 and G_2; $x \in V(G_1)$; R; h
Step 1: Find pair of initial seeds $s_1 \in V(G_1)$ and $s_1' \in V(G_2)$ so that s_1 is adjacent to x in G_1 .
Generate h-hop neighborhoods around initial seeds (be sure VOI is in first neighborhood).
Find more seeds if possible.
Step 2: Record C_x', the set of non-seed vertices in second h-hop neighborhood.

VN via LNM algorithm

Input: Graphs G_1 and G_2; $x \in V(G_1)$; R; h

Step 1: Find pair of initial seeds $s_1 \in V(G_1)$ and $s_1' \in V(G_2)$ so that s_1 is adjacent to x in G_1 .

Generate h-hop neighborhoods around initial seeds (be sure VOI is in first neighborhood).

Find more seeds if possible.

Step 2: Record C_x', the set of non-seed vertices in second h-hop neighborhood.

Step 3: Run SGM algorithm (modified from [FAP12]) for matching the two neighborhoods generated by initial seeds R times. Set P to be the average of all the matchings.

VN via LNM algorithm

Input: Graphs G_1 and G_2; $x \in V(G_1)$; R; h
Step 1: Find pair of initial seeds $s_1 \in V(G_1)$ and $s_1' \in V(G_2)$ so that s_1 is adjacent to x in G_1 .
Generate h-hop neighborhoods around initial seeds (be sure VOI is in first neighborhood).
Find more seeds if possible.
Step 2: Record C_x', the set of non-seed vertices in second h-hop neighborhood.
Step 3: Run SGM algorithm (modified from [FAP12]) for matching the two neighborhoods generated by initial seeds R times. Set P to be the average of all the matchings.
Step 4: $P[x,]$ is the row of probabilities. Top nominate for x' is in $\arg\max_{v \in C_x'} P[x,]$.

Simulations

- Repeatedly generate pairs of graphs from a .6-correlated SBM with probability matrix

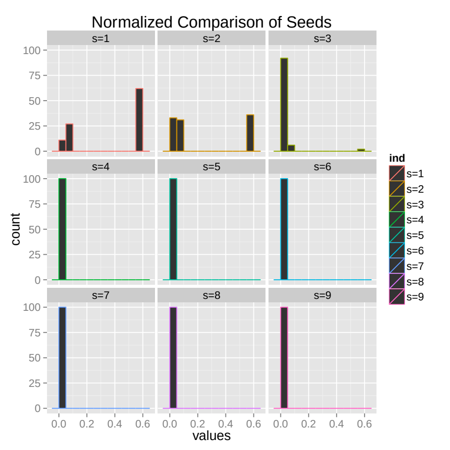
$$\Lambda = \begin{bmatrix} 0.7 & 0.3 & 0.4 \\ 0.3 & 0.7 & 0.3 \\ 0.4 & 0.3 & 0.7 \end{bmatrix}.$$

- Select VOI.
- Steps 1-4.
- Plot normalized rank of x'

$$\frac{\text{rank}(x') - 1}{|C'_x| - 1}$$

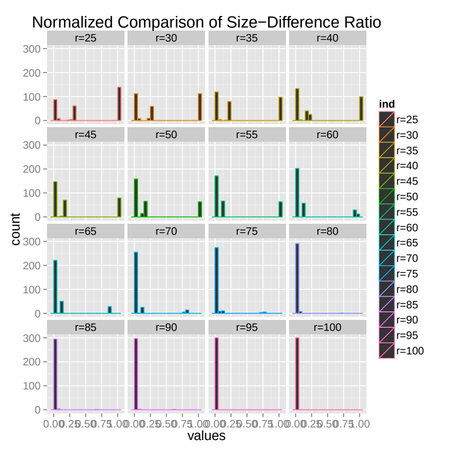
  in histogram.
- NOTE: 0, 0.5, and 1 imply that x' was first, half-way down, and last in the nomination list, respectively.

Effects of number of seeds



Effect of number of seeds on VOI nomination list, using graphs with 300 vertices and 1 VOI. As the number of seeds increases, the location of x' in the nomination list decreases.

Effects of size discrepencies between graphs

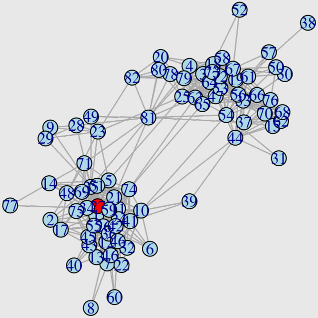


The larger graph has 100 vertices per block, and the smaller graph has r vertices per block. The smaller graph is 0.6 correlated with an induced subgraph of the larger one. Larger graph has $3(100 - r)$ "junk" vertices: As the number of "junk" vertices decreases, the location of x' in the nomination list decreases.

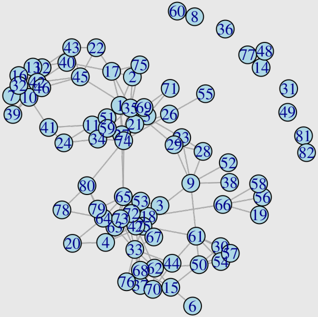
Real Data Experiments

- Core of High School Facebook and Friendship Survey Networks.
- Twitter and Instagram Networks.

High School and Facebook Networks [MFB15]

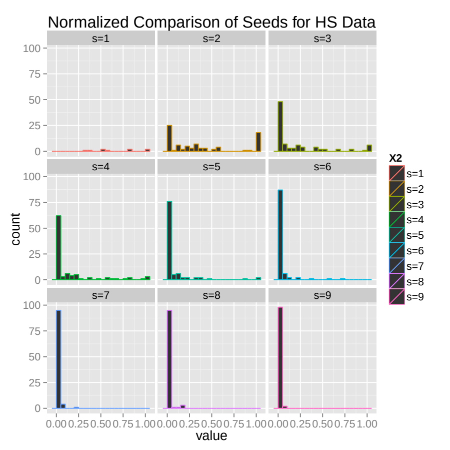


(a) Core of Facebook Friendship Network



(b) Core of Survey-Based Friendship Network

Example of VN-LNM for HS data with VOI 27



In this example, we are stochastically better than uniform by 3 seeds.

Twitter and Instagram Networks

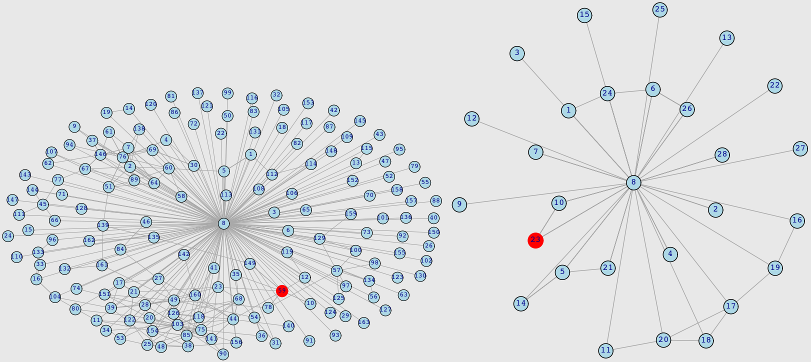
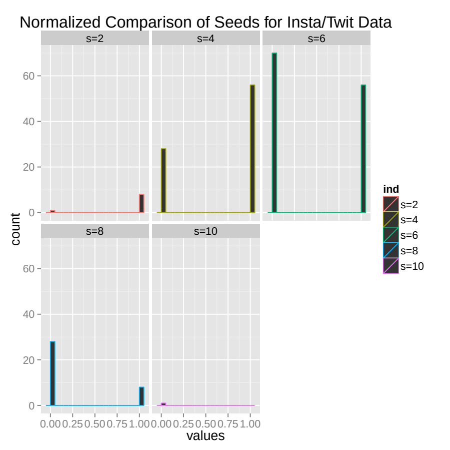


Figure: Twitter network on left and Instagram network on right.

Seeds Benefit in Instagram/Twitter Matching



Fixed VOI and fixed seed-set of size 10. For every subset of size $s \in \{2, 4, 6, 8, 10\}$ we run steps 2-4 and record the normalized rank.

Evidence Suggests...

- As the number of seeds increases, often the proportion of times the true match maps to the VOI increases (i.e. the minimum k required to obtain true match degreases).
- When the vertex sets have differing sizes the matching becomes more difficult.
- The presence of "junk" vertices complicates the problem.

Future Work

- Determine bounds on how much “junk” can be added in ρ-SBM case and still guarantee, at least asymptotically, that we will match the core vertices correctly.
- Explore how choice of seeds can be made (i.e. what makes a good seed).
- What happens when ρ is different based on block structure?
- Extend this work to finding multiple VOI across multiple networks simultaneously.

Acknowledgements

Thank you to the XDATA and SIMPLEX programs of the Defense Advanced Research Projects Agency and to the Acheson J. Duncan Fund for the Advancement of Research in Statistics.

References

[FAP12] D.E. Fishkind, S. Adali, and C.E. Priebe, *Seeded graph matching*, arXiv:1209.0367 (2012).

[MFB15] R. Mastrandrea, J. Fournet, and A. Barrat, *Contact patterns in a high school: a comparison between data collected using wearable sensors, contact diaries and friendship surveys*, PLoS ONE (2015).