Vertex Nomination Via Local Neighborhood Matching

Heather G. Patsolic

Johns Hopkins University

May 17, 2017

Publications

H.G. Patsolic, V. Lyzinski, C.E. Priebe, and Y. Park,
"Vertex Nomination via Local Neighborhood Matching,"
arXiv:1705.00674,
2017

D.E. Fishkind, S. Adali, H.G. Patsolic, L. Meng, V. Lyzinski, and C.E. Priebe,
"Seeded Graph Matching,"
arXiv:1209:0367,
2017.

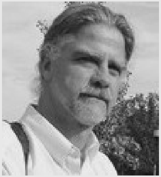
R. Mastrandrea, J. Fournet, and A. Barrat
"Contact patterns in a high school: a comparison between data collected using wearable sensors, contact diaries and friendship surveys,"
PLoS ONE,
2015.

Links to relevant papers and code and data for all simulations and experiments can be found at:

`http://www.cis.jhu.edu/~parky/XDATA/SGM/vn.html`.

Youngser Park

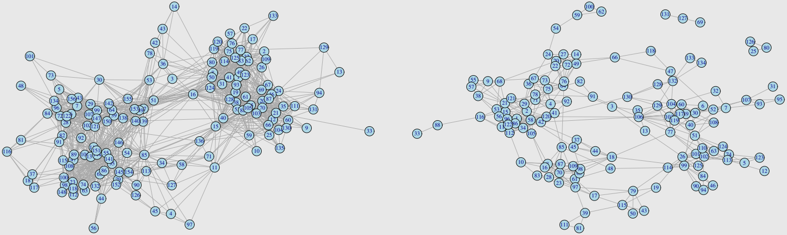
Vince Lyzinski



Carey E. Priebe

Problem Formulation

- Two large networks on overlapping, non-identical vertex sets.
- Seed vertices for which correspondences are known.
- There is a vertex of interest (VOI) in one network we'd like to identify in the other.
- Goal: Find vertex corresponding to VOI in the other network.



(a) Facebook Network

(b) High School Survey Network

Figure: Data obtained from [3].

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.

Course of Action

- Localize the problem
- Localize the problem
- Apply graph matching techniques [2]
- Nominate potential matches to the VOI

Viewing the Facebook graph with VOI at center

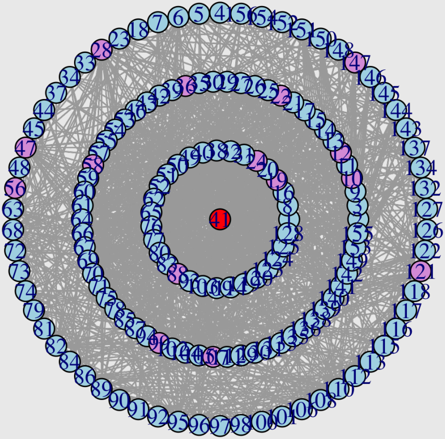


Figure: Facebook graph centered at VOI $x = 41$, with h-hop neighborhoods in concentric circles about x.

Creating Local Seed Set S_x $(h = 1)$

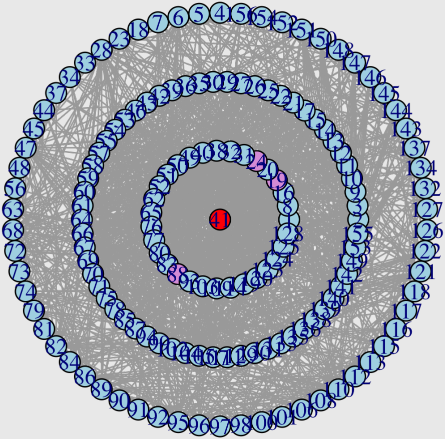
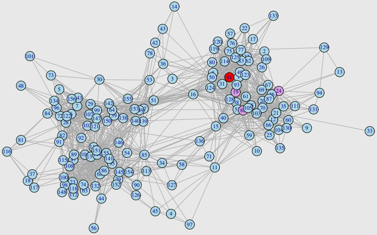
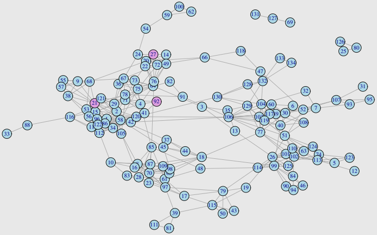


Figure: Local seed set $S_x = \{19, 24, 88\}$ created in Facebook network choosing seeds within a 1-hop neighborhood of the VOI.

Creating Local Seed Set S'_x

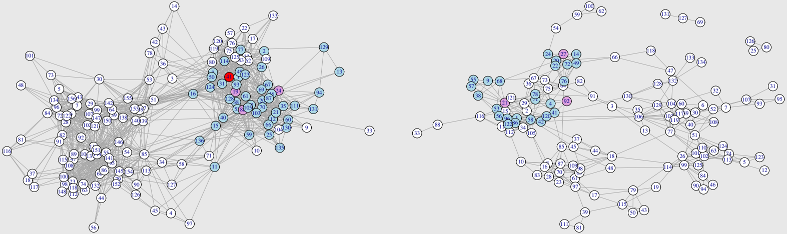


(a) Facebook Network with local seed set $S_x = \{19, 24, 88\}$



(b) High School Survey Network with corresponding local seed set $S'_x = \{21, 27, 92\}$

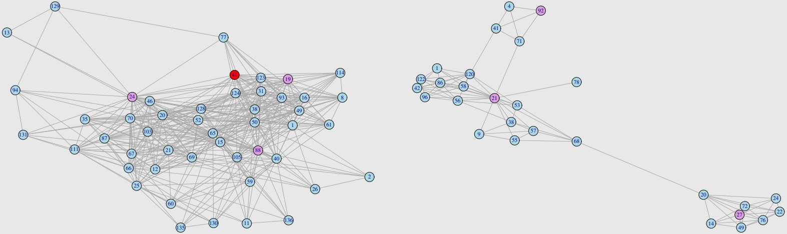
Creating Local Neighborhoods of S_x and S'_x ($\ell = 2$)



(a) Facebook Network with seeds $S_x = \{19, 24, 88\}$

(b) High School Survey Network $S'_x = \{21, 27, 92\}$

Creating Local Neighborhoods of S_x and S'_x ($\ell = 2$)



(a) Facebook Network with seeds $S_x = \{19, 24, 88\}$

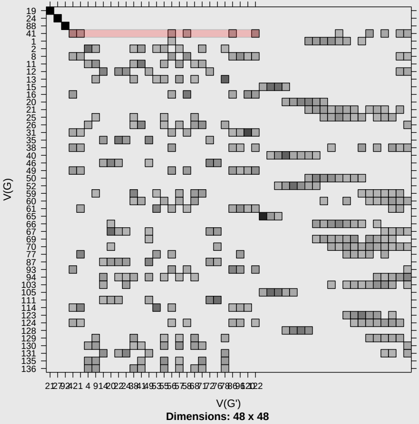
(b) High School Survey Network $S'_x = \{21, 27, 92\}$

Candidate set of vertices is $C'_x = \{1, 4, 9, 14, 20, 22, 24, 38, 41, 42, 49, 53, 55, 56, 57, 58, 68, 71, 72, 76, 78, 86, 96, 120, 122\}$, and $|C'_x| = 25$.

Course of Action

- Localize the problem
- Apply graph matching techniques [2]
- Nominate potential matches to the VOI

Applying Soft Seeded Graph Matching (SoftSGM) [2]

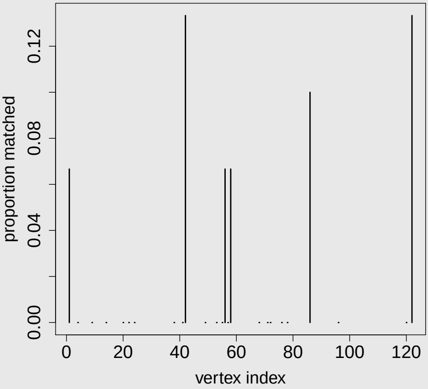


- Apply Seeded Graph Matching (SGM) algorithm [2] repeatedly (R times) and average over solutions.
- Obtain matrix D so that element i, j represents the proportion of times vertex j in G' mapped to vertex i in G.

Course of Action

- Localize the problem
- Apply graph matching techniques [2]
- Nominate potential matches to the VOI

Creating Nomination List



- The most likely nominate for the VOI is in $\arg\max_{v \in C'_x} D[x, v]$.
- Nomination list for the VOI, x, is the list of vertices in C'_x ordered from highest to lowest value in $D[x,]$
- $\Phi_x = (\{42, 122\}, 86, \{1, 55, 57\}, \ldots)$

VNmatch [1]

```
**Input**: Graphs: $G = (V, E)$, $G' = (V', E')$, Seeds/Seeding: $S \leftrightarrow S'$, VOI: $x \in V$, Limits: $h$, $\ell$, Restarts: $R$
*Step 1*: Find seeds within $h$-path of VOI: $S_x$ and $S'_x$
*Step 2*: Create induced subgraphs of $G$ and $G'$ generated by vertices within $\ell$-path of $S_x$ and $S'_x$— Denote by $A$ and $B$ the corresponding adjacency matrices
*Step 3*: Run SoftSGM with $R$ restarts to get matrix $D$
*Step 4*: Create a nomination vector $\Phi_x$ based on the proportion of times $u \in V'$ is matched to $x$ according to the vector $D(x, :)$.
**return** $\Phi_x$ the nomination vector of likely matches to $x$.
```

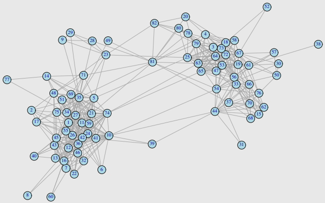
Notation Used in Examples

Let C'_x denote the set of candidate vertices for x, that is, the set of non-seed vertices in the induced subgraph of G', $rank(x')$ denote the expected rank (location) of x' in the nomination list, and define

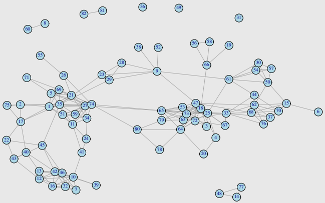
$$\tau(x') = \left(\frac{rank(x') - 1}{|C'_x| - 1}\right). \tag{1}$$

Example: $rank(42) = 1.5$ and $|C'_x| = 25$, so
$\tau(x') = \frac{1.5-1}{25-1} = \frac{1}{48} \approx .021$.

High School and Facebook Networks [3]

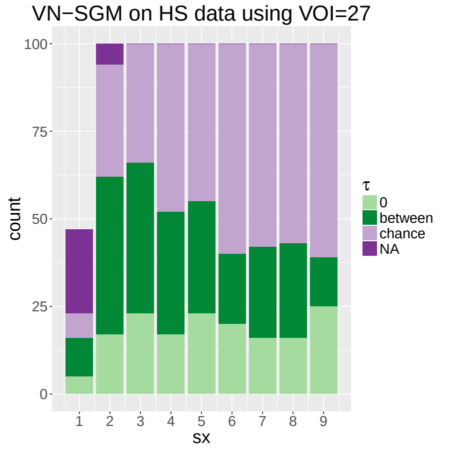


(a) Core of High School Friendship Network based on Facebook

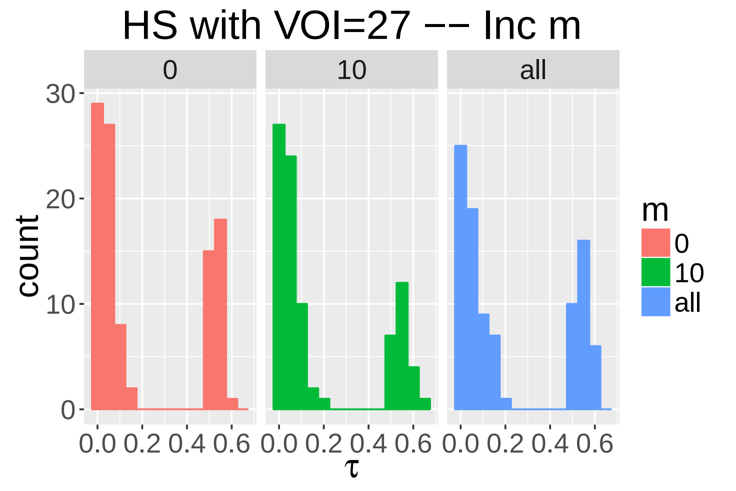


(b) Core of High School Friendship Network based on the Survey

Applying `VNmatch` to HS core networks with VOI 27 (41)



Adding unshared vertices to HS networks



Twitter and Instagram Networks

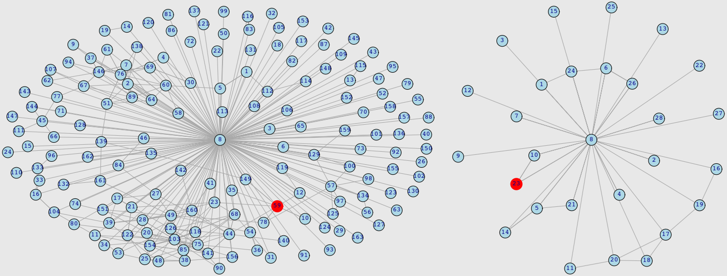


Figure: (left) Twitter network (right) Instagram network

VNmatch applied to Instagram and Twitter Networks

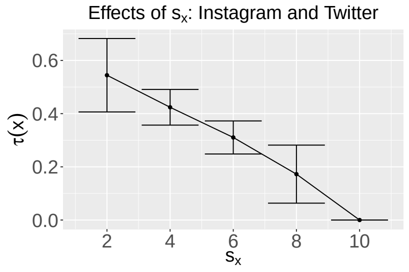


Figure: Fixed VOI and fixed seed-set of size 10. For every subset of size s (even) we run VNmatch algorithm and record the location of the VOI in the nomination list. We plot the average and CI (mean $\pm$ 2*se) for each $s \in \{2, 4, 6, 8, 10\}$.

Future Work

- Explore the effects of unshared vertices and how to address them in more detail.
- Explore how choice of seeds can be made (i.e. what makes a good seed).
- In the SBM setting, what happens when ρ is different based on block structure?

Acknowledgements

Thank you to the XDATA, D3M, and SIMPLEX programs of the Defense Advanced Research Projects Agency, and to the Acheson J. Duncan Fund for the Advancement of Research in Statistics (awards 16-20 and 16-23).

Mathematical Framework for Simulations: ρ-SBM

$G, G' \sim \rho$–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 and vertices i and j in G' has correlation ρ.
- Otherwise, edge presence is conditionally independent across graphs.

Simulations

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

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

- For each ρ, select VOI and s_x seeds uniformly at random.
- Apply `VNmatch` algorithm
- Plot average normalized rank $\tau(x')$ as a function of s_x and ρ.

Effects of correlation and number of seeds

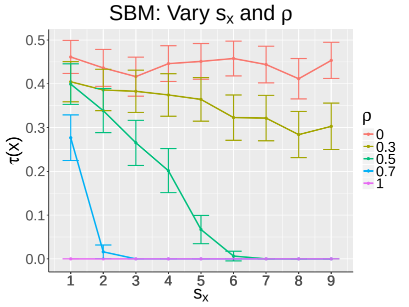


Figure: Demonstration of the effects of correlation between graphs and number of seeds used in matching on performance of the `VNmatch` algorithm.

Effects of size discrepencies between graphs

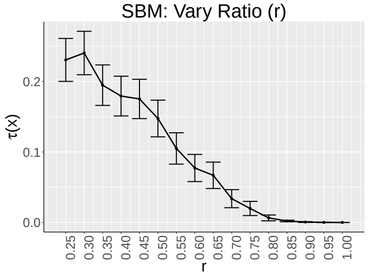


Figure: Letting $r = |V|/|V'|$ denote the ratio between the number of vertices in the smaller graph to the number of vertices in the larger graph, we plot the average value $\tau(x')$ as a function of r using 4 seeds and $\rho = 0.6$..

The General Graph Matching Problem (GMP): Definition

Definition (Graph Matching)

The **graph matching** problem aims to solve the following objective function:

$$\min_{P\in\Pi(\mathfrak{n})} \left\| A - PBP^T \right\|_F^2 = \min_{P\in\Pi(\mathfrak{n})} \| AP - PB \|_F^2 . \qquad (2)$$

Relaxing the General GMP

The GMP relaxes to a convex quadratic program:

$$\min_{D \in \mathcal{D}(\mathfrak{n})} \|AD - DB\|_F^2 . \tag{3}$$

An alternative formulation (no longer equivalent objective value) is the indefinite, quadratic formulation:

$$\max_{D \in \mathcal{D}(\mathfrak{n})} \text{trace}(A^T D B D^T). \tag{4}$$

The doubly stochastic solution is then projected back onto $\Pi(\mathfrak{n})$, giving a solution to the GMP.

Adding Seeds to Graph Matching

Given seed-sets S and S' with seeding $S \leftrightarrow S'$. WLOG: $S = S' = \{1, \ldots, s\}$.

Definition (Seeded Graph Matching (SGM))

The **seeded graph matching** problem aims to solve the following objective function:

$$\min_{P \in \Pi(n)} \|A(I \oplus P) - (I \oplus P)B\|_F^2 . \tag{5}$$

where I denotes the s-by-s identity matrix and $n = \mathfrak{n} - s$

Relaxing the Seeded Graph Matching (SGM) Problem

Similarly we relax the SGMP to:

$$\min_{D \in \mathcal{D}(n)} \|A(I \oplus D) - (I \oplus D)B\|_F^2 \, . \tag{6}$$

An alternative formulation (no longer equivalent objective value) is the indefinite, quadratic formulation:

$$\max_{D \in \mathcal{D}(n)} \operatorname{trace}(A^T(I \oplus D)B(I \oplus D^T)). \tag{7}$$

The doubly stochastic solution is then projected back onto $\Pi(n)$, giving a solution to the SGMP.
Tools for solving **??**: Frank-Wolfe and Hungarian algorithms.

Maximizing $f(D) = \text{trace}(A^T(I \oplus D)B(I \oplus D^T))$: Frank-Wolfe

1. Initialize $D^{(0)}$
2. Compute $\nabla_D f(D)|_{D^{(i)}}$
3. Compute $Q \in \mathcal{D}(n)$ to maximize $\text{trace}(Q^T \nabla f(D^{(i)}))$ via the Hungarian Algorithm
4. Compute step size $\alpha \in [0, 1]$ to maximize $f(\alpha D^{(i)} + (1 - \alpha)Q)$
5. Set new iterate $D^{(i+1)} = \alpha D^{(i)} + (1 - \alpha)Q$
6. Continue until maximum number of iterates or stopping tolerance met

Solving the SGMP

1. Relax the problem
2. Solve the relaxation via Frank-Wolfe methodology
3. Project final iterate from Frank-Wolfe back to the permutation matrices

These last two steps consistute the SGM algorithm, which is the FAQ algorithm of [**?**] when $s = 0$.

Demonstrating SGM with ρ-correlated Stochastic Blockmodel (ρ-SBM) Example

$(G, G') \sim \rho - SBM(k, b, \Lambda)$ if

(1) First, G and G' are marginally stochastic blockmodel graphs, $G \sim SBM(k, b, \Lambda)$ and $G' \sim SBM(k, b, \Lambda)$

(i) k is a positive integer representing the number of blocks in each graph,

(ii) $b : V \to \{1, 2, \ldots, k\}$ is a map assigning to each vertex in V a block label,

(iii) $\Lambda \in [0, 1]^{k \times k}$ is a probability matrix such that

$$\mathbb{1}\{\{v, w\} \in E\} \overset{ind.}{\sim} \text{Bernoulli}(\Lambda_{b(v),b(w)}).$$

(2) the Pearson correlation coefficient between $\mathbb{1}\{\{v_i, v_j\} \in E\}$ and $\mathbb{1}\{\{u_i, u_j\} \in E'\}$ is ρ, and edge presence across graphs is otherwise independent.

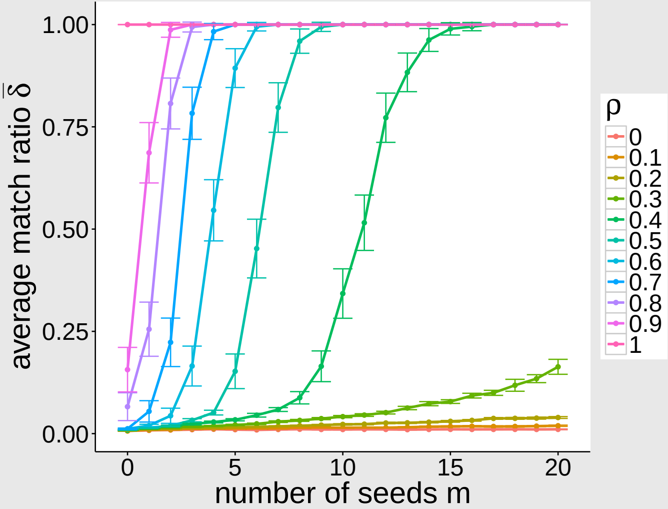
Measuring Performance of SGM: Match Ratio

The *match ratio* is defined to be the fraction of non-seed vertices of G that are correctly matched:

$$\delta := \frac{|\{v \in V \backslash S : \phi(v) = \Psi(v)\}|}{n}. \tag{8}$$

$\overline{\delta}$ is the average proportion of times that a vertex is correctly matched over many simulations.

Measuring Performance of SGM for various s and ρ



Addressing Unshared Vertices

- Unknown how many of the non-seed vertices are shared vertices as opposed to unshared.
- Approach: Add "phantom"isolated vertices to the smaller graph. WLOG: A smaller.
- Consider SGM for $2A - \mathbb{1}\mathbb{1}^T \oplus [0]$ and $2B - \mathbb{1}\mathbb{1}^T$.
- This forces stronger penalty for edge mismatches when all vertices are in the graphs and weaker penalty for edge mismatches when some vertices are "phantom" vertices.