(Robust) Seeded graph matching

Vince Lyzinski
Human Language Technology Center of Excellence
Dept. of Applied Mathematics and Statistics
Johns Hopkins University
Baltimore, MD USA

August 5, 2014

Collaborators

Carey E. Priebe, Joshua Vogelstein, Donniell Fishkind, Daniel Sussman, Avanti Athreya, Youngser Park, Sancar Adali

- Given two graphs, the Graph matching problem (GMP) seeks to find an alignment of the vertex sets of the two graphs that best preserves the connectivity structure of the graphs

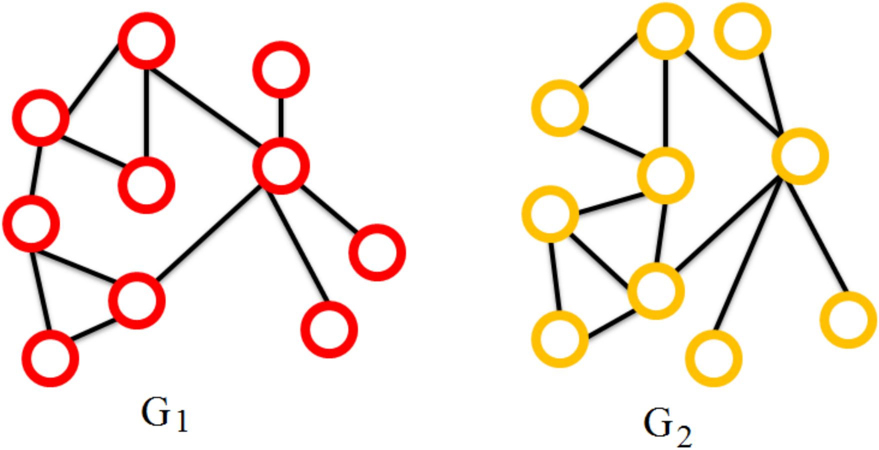
International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 18, No. 3 (2004) 265–298

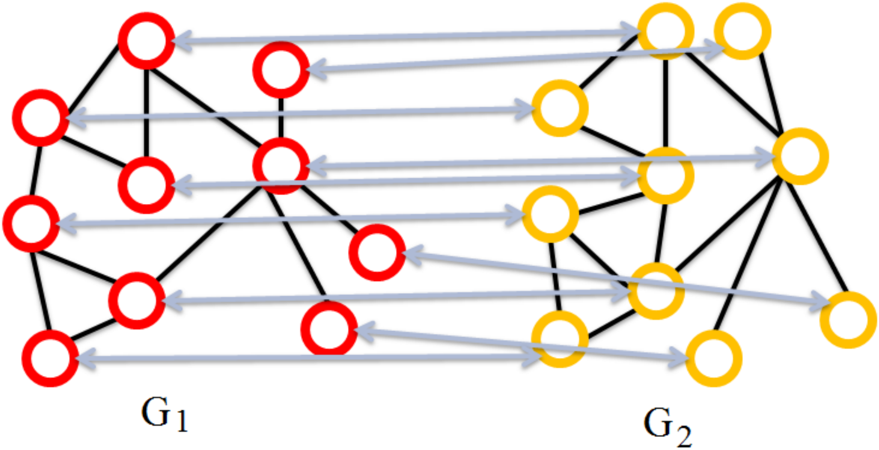
World Scientific
www.worldscientific.com

THIRTY YEARS OF GRAPH MATCHING IN PATTERN RECOGNITION

D. CONTE[*,‡], P. FOGGIA[†,§], C. SANSONE[†,¶] and M. VENTO[*,‖]

[*] *Dipartimento di Ingegneria dell'Informazione e di Ingegneria Elettrica, Università di Salerno – Via P.te Don Melillo,1 I-84084, Fisciano (SA), Italy*
[†] *Dipartimento di Informatica e Sistemistica, Università di Napoli "Federico II" – Via Claudio, 21 I-80125 Napoli, Italy*
[‡] *dconte@unisa.it*


G_1
G_2


G_1
G_2

- Given two graphs, the Graph matching problem (GMP) seeks to find an alignment of the vertex sets of the two graphs that best preserves the connectivity structure of the graphs
- In the seeded graph matching problem, we are further provided subsets of the vertices (Seeds)

  $$S_1 \subset V_1, \ S_2 \subset V_2$$

  and a seeding function

  $$\phi \subset S_1 \times S_2, \quad \text{i.e. } (u, v) \in \phi \Rightarrow \text{ u and v are "matched"}$$

  Goal: Leverage ϕ to match the remaining vertices

- Our proceudre: The Joint Optimization of Fidelity and Commensurabilty (JOFC) graph matching algorithm of Lyzinski et al. (2014)
- Adapted from the manifold matching methodologies of Priebe et al. (2013)
- Is **robust**
- Briefly, the JOFC algorithm proceeds as follows:
  1. Jointly embed the seeded vertices into a common Euclidean space
  2. Separately out-of-sample embed the nonseeded vertices
  3. Match the embedded nonseeded vertices—this is equivalent to solving the generalized assignment problem

- We begin with Δ_1 and Δ_2, dissimilarity representations of the graphs G_1 and G_2 resp.
- Let

$$X_1^{(1)}, X_2^{(1)}, \ldots, X_{s_1}^{(1)} \text{ be the embedded seeded vertices of } G_1,$$

and

$$X_1^{(2)}, X_2^{(2)}, \ldots, X_{s_2}^{(2)} \text{ the embedded seeded vertices of } G_2$$

- We want the embedding to simultaneously preserve:
  1. The within graph dissimilarities amongst the seeded vertices (fidelity)

  1b. The across graph relationship provided by the seeding between non-matched seeded vertices (separability)

  2. The across graph relationship given by the seeding (commensurability)

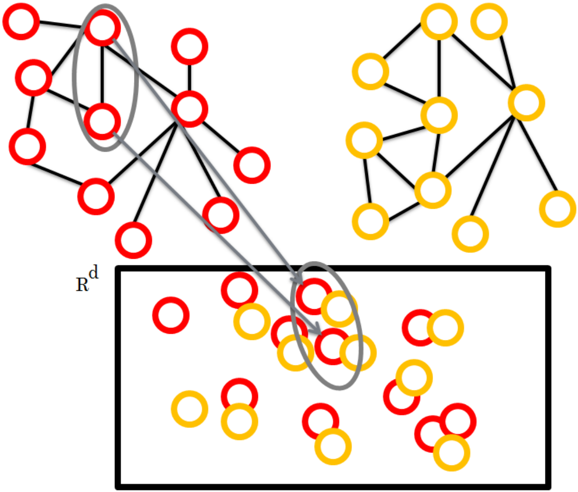
- We want the embedding to simultaneously preserve:
  1. The within graph dissimilarities amongst the seeded vertices i.e. we wish to minimize the within graph squared fidelity error of the embedding given by

$$\varepsilon^2_{F_1} := \sum_{i,j} \left(d(X_i^{(1)}, X_j^{(1)}) - \Delta_1(i,j) \right)^2$$

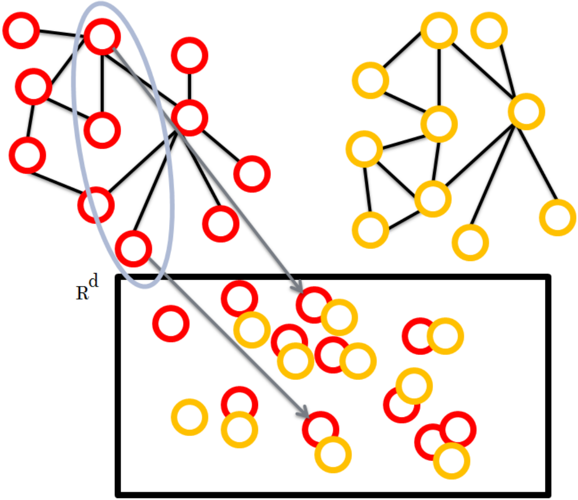
and

$$\varepsilon^2_{F_2} := \sum_{i,j} \left(d(X_i^{(2)}, X_j^{(2)}) - \Delta_2(i,j) \right)^2$$

Minimize the squared fidelity error of the embedding



Minimize the squared fidelity error of the embedding



- We want the embedding to simultaneously preserve:
  1b. The dissimilarities amongst the non-matched seeded vertices across graphs
  i.e. we wish to minimize the across graph squared separability error of the embedding given by

$$\varepsilon_S^2 := \sum_{\substack{\text{i,j s.t. i and j are non-matched} \\ \text{seeded vertices}}} \left(d(X_i^{(1)}, X_j^{(2)}) - \delta(i,j) \right)^2$$

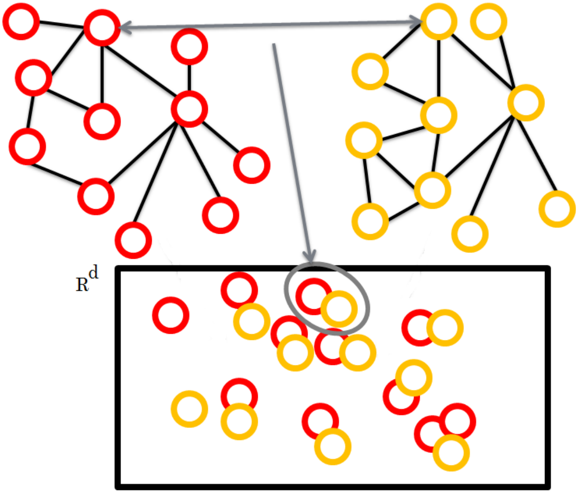
  δ here is an unknown across graph dissimilarity, and must be imputed or treated as missing data

- We want the embedding to simultaneously preserve:
  2. The across graph relationship given by the seeding i.e. we wish to minimize the within graph squared commensurability error of the embedding given by

$$\varepsilon_C^2 := \sum_{\substack{\text{i,j s.t. i and j are matched} \\ \text{by the seeding}}} \left(d(X_i^{(1)}, X_j^{(2)}) - \delta(i,j) \right)^2$$

     Again δ here is an unknown across graph dissimilarity, though if i and j are matched by the seeding, it is reasonable to impute $\delta(i,j) = 0$

2. Minimize the squared commensurability error of the embedding



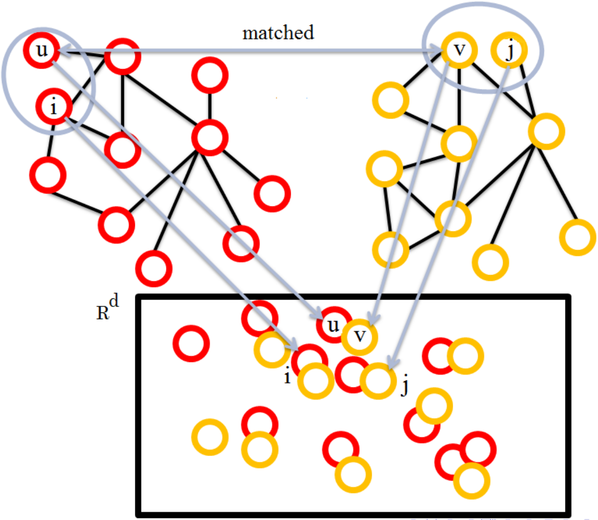
- We embed the seeded vertices using the SMACOF algorithm of de Leeuw (1977)—a version of weighted MDS
- With a judicious choice of weightings here, SMACOF seeks to minimize
$$w(\epsilon_{F_1}^2 + \epsilon_{F_2}^2 + \epsilon_S^2) + (1-w)\epsilon_{C^2}$$
- In our applications, we found best performance with $w = 0.8$; see Adali et al. (2013) for motivation

- We next out-of-sample embed the unseeded vertices of G_1 ($\{Y_i^{(1)}\}$) and of G_2 ($\{Y_i^{(2)}\}$) using the weighted MDS procedure of Tang et al. (2013)
- Through a judicious choice of weightings, we seek to minimize the stress function

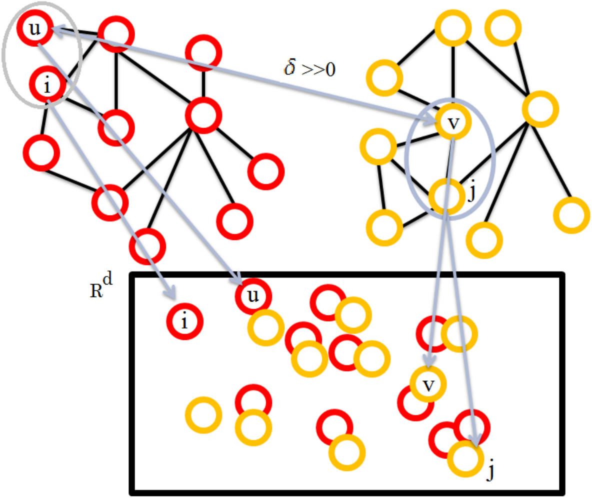
$$\sigma(\mathbf{Y}) = \sum_i \sum_j \left(d(X_i^{(1)}, Y_j^{(1)}) - \Delta^{(1)}(i,j) \right)^2 + \sum_i \sum_j \left(d(X_i^{(2)}, Y_j^{(2)}) - \Delta^{(2)}(i,j) \right)^2$$

  i.e. we wish to preserve the within graph dissimilarities between the seeded and unseeded vertices
- Then match embedded vertices based on Euclidean distance: closer together=more likely to be matched

- Suppose $i \in V(G_1)$ and $j \in V(G_2)$ are unseeded vertices that should be matched to each other



- Suppose $i \in V(G_1)$ and $j \in V(G_2)$ are unseeded vertices that should NOT be matched to each other



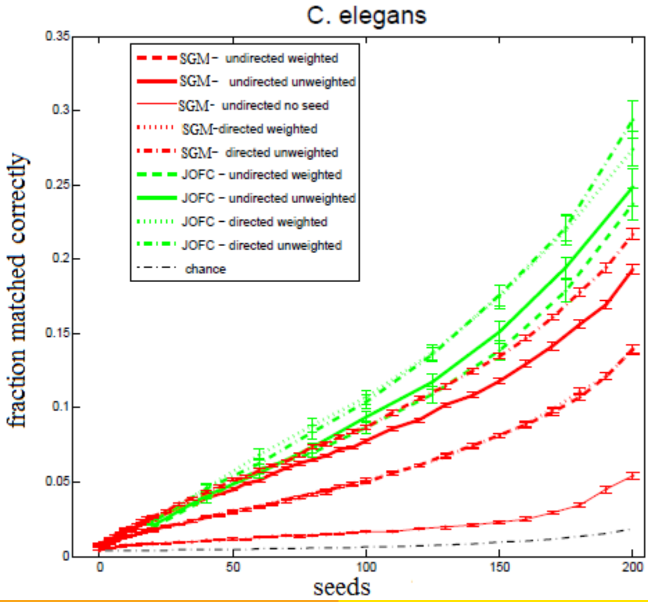
- We then match the unseeded embedded vertices by minimizing

$$\sum_{\substack{i,j \text{ unseeded vertices} \\ \text{matched by M}}} d(Y_i^{(1)}, Y_j^{(2)}).$$

  over all matchings $M \subset V(G_1) \times V(G_2)$ of the unseeded vertices
- This is equivalent to a generalized assignment problem—NP-hard in general, but are good approximation algorithms

Ex Matching C. elegans connectomes

- The C. elegans nervous system is believed to be composed of 302 labelled vertices (279 with synapses to other neurons)
- There are two types of connections between neurons: electrical and chemical
- We match the chemical connectome to the electrical connectome to understand the extent to which connectivity structure alone is enough to identify neurons across the connectomes


C. elegans
SGM- undirected weighted
SGM- undirected unweighted
SGM- undirected no seed
SGM-directed weighted
SGM- directed unweighted
JOFC – undirected weighted
JOFC – undirected unweighted
JOFC – directed weighted
JOFC – directed unweighted
chance
fraction matched correctly
seeds
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0
50
100
150
200

- This procedure is flexible enough to handle:
  1. $|V(G_1)| \neq |V(G_2)|$—allow for many–to–many, many–to–one or many–to–none matchings in the embedded graphs
  2. Weighted directed graphs—use a dissimilarity that allows for weightedness and directedness in the graphs for the embedding
  3. The case of "soft" seeds—suitably weight the MDS embeddings
- Is easily modified for vertex classification
- Working on scaling the SMACOF subroutine—would lead to our algorithm scaling

This work is partially supported by a National Security Science and Engineering Faculty Fellowship (NSSEFF), Johns Hopkins University Human Language Technology Center of Excellence (JHU HLT COE), and the XDATA program of the Defense Advanced Research Projects Agency (DARPA) administered through Air Force Research Laboratory contract FA8750-12-2-0303.

References

S. Adali, C. E. Priebe. Fidelity-commensurability tradeoff in joint embedding of disparate dissimilarities. *arXiv preprint, arXiv:1306.1977* (2013).

D. Conte, P. Foggia, C. Sansone, and M. Vento. Thirty years of graph matching in pattern recognition. *International journal of pattern recognition and artificial intelligence*, **18(03)**, (2004).

J. de Leeuw. Applications of convex analysis to multidimensional scaling. *Recent developments in statistics: Proc. European Meeting of Statisticians, Grenoble*, (1977).

V. Lyzinski, S. Adali, J.T. Vogelstein, Y. Park, and C.E. Priebe. Seeded graph matching via joint optimization of fidelity and commensurability. *arXiv preprint, arXiv:1401.3813* (2014).

V. Lyzinski, D.E. Fishkind, and C.E. Priebe. Seeded graph matching for correlated Erdős-Rényi graphs. *JMLR*, accepted (2014).

C.E. Priebe, D.J. Marchette, Z. Ma, and S. Adali. Manifold matching: joint optimization of fidelity and commensurability. *Brazilian Journal of Probability and Statistics*, **27(3):377400** (2013).

M. Tang, Y. Park, and C. E. Priebe. Out-of-sample extension for latent position graphs. *arXiv preprint, arXiv:1305.4893*, (2013).