Spectral Clustering for Divide-and-Conquer Graph Matching

Carey E. Priebe

Department of Applied Mathematics & Statistics
Johns Hopkins University, Baltimore, MD, USA

January 13-14, 2015
Abstract

We present a parallelized bijective graph matching algorithm that leverages seeds and is designed to match very large graphs. Our algorithm combines spectral graph embedding with existing state-of-the-art seeded graph matching procedures. We justify our approach by proving that modestly correlated, large stochastic block model random graphs are correctly matched utilizing very few seeds through our divide-and-conquer procedure. We also demonstrate the effectiveness of our approach in matching very large graphs in simulated and real data examples.


V. Lyzinski
Given two graphs, $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, the Graph Matching Problem (GMP) seeks an alignment between the vertex sets $V_1$ and $V_2$ that best preserves structure across the graphs. In bijective graph matching, we further assume $|V_1| = |V_2| = n$, and the alignment sought by GMP is a bijection between $V_1$ and $V_2$.

**Graph Matching Problem**

Find a bijection $\psi : V_1 \rightarrow V_2$ minimizing the quantity

$$\left| \left\{ (i, j) \in V_1 \times V_1 \text{ s.t. } [i \sim_{G_1} j, \psi(i) \sim_{G_2} \psi(j)] \text{ or } [i \sim_{G_1} j, \psi(i) \nsim_{G_2} \psi(j)] \right\} \right|,$$

i.e. the problem seeks to minimize the number of edge disagreements between $G_2$ and “$\psi(G_1)$”. Equivalently stated, if $A$ and $B$ are the respective adjacency matrices of $G_1$ and $G_2$, then this problem seeks to minimize $\|A - PBP^T\|_F^2$, over all permutation matrices $P \in \Pi(n) := \{n \times n \text{ permutation matrices}\}$, with $\| \cdot \|_F$ the matrix Frobenius norm.
In the seeded graph matching problem (SGMP), we further assume the presence of a latent alignment $\phi$ between the vertex sets of $G_1$ and $G_2$. Our task is to then efficiently leverage the information in a partial observation of the latent alignment, i.e. a *seeding*, to estimate the remaining latent alignment.

**Seeded Graph Matching Problem**

Given subsets of the vertices $S_1 \subset V_1$ and $S_2 \subset V_2$ called *seeds* with $|S_1| = |S_2| = s$ and a bijective seeding function $\phi_S : S_1 \to S_2$, the task is to use $\phi_S$ to estimate $\phi$ by finding the bijection extending $\phi_S$ which minimizes (1).
Divide-and-Conquer Seeded Graph Matching

\[ \Omega(C_{i,1}, G_1) \xlongequal{SGM} \Omega(C_{i,2}, G_2) \Rightarrow \psi^{(i)} \]

\[ \psi = \bigoplus_{i=1}^{k} \psi^{(i)} \]
Theorems

Theorem 1: Perfect Clustering

[EJS2014]

Theorem 2: Seeded Graph Matching

[JMLR2014]

Theorem 3: Subspace Alignment

[PARCO2015]


Fraction of unseeded vertices correctly matched across two $K = 900$ block, $\bar{n} = 30 \cdot \bar{1}$, $d = 10$ dimensional $\rho$-correlated SBM's with $s$ seeds drawn uniformly at random from the 27000 vertices.
The fraction of the unseeded vertices correctly matched for graphs 8 and 29 (within-subject) and for graphs 1 and 8 (across-subject). For the 8–29 pair, \( n = 20,541, d = 30 \). For the 1–8 pair, \( n = 18,694, d = 30 \), we cluster using \( k \)-means, reclustering any clusters of size \( \geq 800 \). We plot the fraction of the vertices correctly matched in each of the two experiments for number of seeds \( s = 200, 1000, 2000, \) and 5000.
Yogi Berra:

“In theory there is no difference between theory and practice. In practice, there is.”
Leopold Kronecker to Hermann von Helmholtz (1888):

“The wealth of your practical experience with sane and interesting problems will give to mathematics a new direction and a new impetus.”