Link Prediction in Undirected Networks

A Probabilistic Foundation

Kevin Miller
Supervised by: Dr. Jeffrey Humpherys

March 4, 2017

Motivation

Graphs (Networks) are collections of nodes (vertices) and edges (connections) that can be used to model relationships.

- Social Networks (Facebook, Twitter, LinkedIn, ego-nets)
- Protein-protein interaction networks
- Telecommunication Networks
- Buying/Selling Networks (Amazon, Ebay, etc.)

Motivation

Problems:

- Community Detection, Clustering, Partitioning
- Centrality Measures
- Link Prediction
- Graph Drawing/Visualization
- Diffusion Analysis

We look at Link Prediction today, using an Effective Resistance based metric

Graph Basics

Graph

Graph $G(V,E)$ is a set of n nodes (vertices) $V = \{1, 2, \dots, n\}$, with pairs of nodes connected by edges (links) in the set E.

Adjacency Matrix

$$A_{ij} = \begin{cases} 1 & \text{if there exists an edge in } E \text{ from node } i \text{ to node } j, \\ 0 & \text{otherwise.} \end{cases}$$

Note we can replace 1 by weight w_e for the weight of edge $e = (i,j)$.

Graph Basics (cont.)

Degree Matrix

$$D = diag(d_1, d_2, \ldots, d_n)$$

$$where\ d_i = deg(i) = \#\ of\ nodes\ that\ node\ i\ connects\ to.$$

Graph Laplacian

Given the adjacency matrix A and degree matrix D,

$$L = D - A$$

Other Graph Laplacians:

- $L_{rw} = I - D^{-1}A$
- $L_{sym} = I - D^{-1/2}AD^{-1/2}$

Directed vs Undirected

Edges $\implies$ relationships or flow of information.

Not all relationships in the world are mutual, or bidirectional.

Undirected Graph

A graph G is **undirected** *if all of the connections are bidirectional. This is equivalent to A and L being symmetric.*

Otherwise, the graph is **directed** *(digraph), with A and L not symmetric.*

Note: $\lambda = 0 \in \sigma(L)$, with eigenvector $\boldsymbol{e} = \mathbb{1}$.

Undirected Example

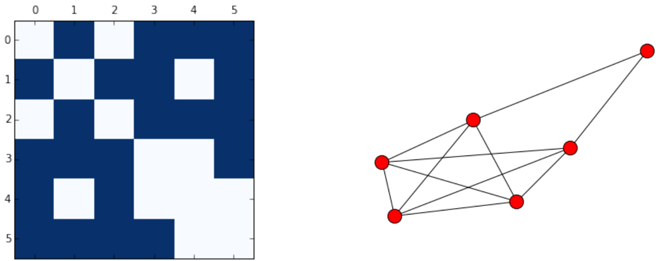


Figure: Undirected Adjacency Matrix and Corresponding Graph

Focus Today : Undirected Networks

We focus only on *undirected* networks, so we have symmetry.

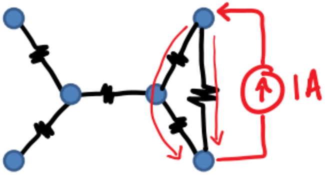
Link Prediction Problem

Given an observed, undirected network $G(V, E)$, what is the most likely *unobserved* edge $e \notin E$ that should be in E, or is likely to be in E in the future?

Problems

- *ill-posed problem*
- *how to measure quality of link prediction?*
- *complex nature of networks, underlying dynamics*

Effective Resistances



Effective Resistance

The effective resistance between nodes $i, j \in V$ is the energy dissipation when a unit current is injected at node i and removed at node j. It can be calculated as the potential difference

$$Reff(i, j) = v(i) - v(j)$$

photo credit: Nikhil Srivastava, Graph Sparsification I: Sparsification via Effective Resistances

Effective Resistances

Model our network as an electrical resistor network:
It can be shown using Kirchoff and Ohm's laws that $Reff(i,j)$ can be found via:

$$Reff(i,j) = (\boldsymbol{e}_i - \boldsymbol{e}_j)^T L^\dagger (\boldsymbol{e}_i - \boldsymbol{e}_j) = L^\dagger_{ii} - 2L^\dagger_{ij} + L^\dagger_{jj}$$

- $L^\dagger$: Moore-Penrose Pseudoinverse of the Graph Laplacian (symmetric)
- $\boldsymbol{e}_i, \boldsymbol{e}_j$: i^{th} and j^{th} standard $\mathbb{R}^n$ basis vectors

Sparsification via Effective Resistances, Daniel Spielman and Nikhil Srivastava

Spielman, Srivastava (2009)

Sparsify dense graphs via random sampling of edges based on the effective resistances across edges.

$$\text{dense } G(V,E) \quad \underset{Reff_G()}{\Longrightarrow} \quad \text{sparse } H(V,E_s)$$

Properties of Sparsification via Effective Resistances

Sparsified graph H retains certain "spectral" properties of G:

- eigenvalues and eigenvectors are "close"
- graph cuts
- clustering

If L_G, L_H are the corresponding Graph Laplacians of G and H, respectively:

$$(1-\epsilon)\boldsymbol{x}^T L_G \boldsymbol{x} \leq \boldsymbol{x}^T L_H \boldsymbol{x} \leq (1+\epsilon)\boldsymbol{x}^T L_G \boldsymbol{x}$$

$\forall \boldsymbol{x} \in \mathbb{R}^n$ with high probability.

Connecting Sparsification and Link Prediction

Sparsification:

$$\text{dense } G(V,E) \implies \text{sparse } H(V,E_s)$$

Link Prediction:

$$\text{“sparse” } H(V,E_s) \implies \text{“dense” } G(V,E)$$

With $e = (i, j) \in E$, we have that

$$Reff(e) : E \to [0, \infty)$$

defines a metric on the edge set, E.

- Effective Resistances $\implies$ "distance"
- more short paths $\implies$ lower Reff() $\implies$ "closer" electrically

Link Prediction via Effective Resistances

With $e = (i, j) \in E$, we have that

$$Reff(e) : E \to [0, \infty)$$

defines a metric on the edge set, E.

- Effective Resistances $\implies$ "distance"
- more short paths $\implies$ lower Reff() $\implies$ "closer" electrically

Extend this metric to **all pairs of nodes.**

Link Prediction Routine

Given an observed, undirected graph $G(V, E)$, we predict the link $\hat{e} \notin E$ s.t.

$$\hat{e} = \underset{e \notin E}{\operatorname{argmin}}\, Reff(e) = \underset{(i,j) \notin E}{\operatorname{argmin}}\, L^{\dagger}_{ii} - 2L^{\dagger}_{ij} + L^{\dagger}_{jj}$$

where $L^{\dagger}$ is the Moore-Penrose Pseudoinverse of the Graph Laplacian.

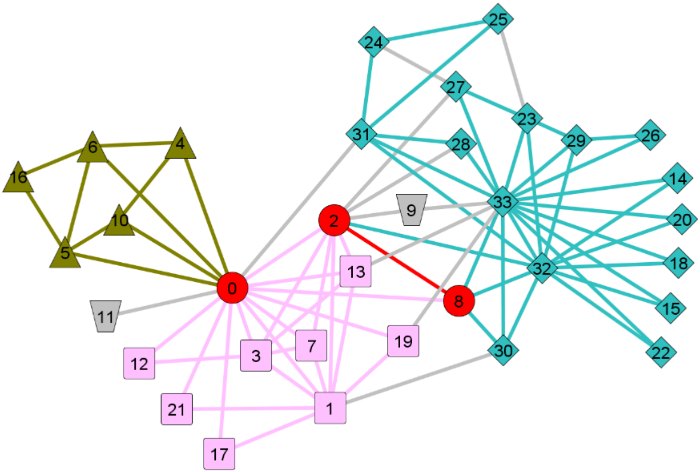


Figure: Zachary's Karate Club Network

Justification for this Method?

Questions:

- Empirically good, but justified?
- In what sense is this predicted link, "the best" or "most likely"?
- Different metrics, different results? Which is best?
- Computationally efficient?

Justification for this Method?

We show that Link Prediction via Effective Resistances yields the "most likely" link in a probabilistic sense, when we view the observed graph as a draw from the probability distribution across edges as defined for Sparsification via Effective Resistances.

Formalization of Link Prediction

Consider an observed, undirected graph $G_o(V, E_o)$ with edge weights $\{w_e\}_{e \in E_o}$ then we define:

- *plus-one graph* = a graph $G_1(V, E_1)$ s.t. $E_1 = E_o \cup \{e_1\}$, with $(e_1 \notin E_o)$
- $\mathbb{G} = \{G_1(V, E_1) : G_1$ is a plus-one graph of $G_o(V, E_o)\}$
- $\mathbb{E} = \{E_1 : E_1$ is a plus-one edge set of $E_o\}$
- $Reff_{E_o}(e)$ = effective resistance of the edge e in the edge set E_o

Probabilistic Foundation for Link Prediction via Effective Resistances

Theorem

Given an undirected, observed graph $G_o(V, E_o)$ and a prior on all edge weights $\{w_e\}_{e \notin E_o}$, the edge $\hat{e} \notin E_o$ s.t.

$$\hat{e} = \operatorname*{argmin}_{e \notin E_o} w_e Reff_{E_o}(e)$$

then $\hat{G}(V, E_o \cup \{\hat{e}\})$ is most-likely plus-one graph to have produced $G_o(V, E_o)$.

Bibliography

- D. Spielman and N. Srivastava. *Graph Sparsification by Effective Resistances*, 2008. http://arxiv.org/abs/0803.0929. Accessed online
- N. Srivastava. *Graph Sparsification I: Sparsification via Effective Resistances*, 2014. Lecture accessed September, 2015 at https://simons.berkeley.edu/talks/nikhil-srivastava-2014-08-26a
- M.E.J. Newman. *The structure of scientific collaboration networks*. Proceedings of the National Academy of Sciences of the United States of America, January, 2001. 10.1073/pnas.021544898
- R. Lichtenwalter, J. Lussier, and N. Chawla. *New Perspectives and Methods in Link Prediction*, 2010. http://doi.acm.org/10.1145/1835804.1835837
- D. Spielman. *Spectral Graph Theory: Effective Resistance*. Lecture 8, Sep 24, 2012. http://www.cs.yale.edu/homes/spielman/561/2012/lect08-12.pdf