Methods to Determine Node Centrality and Clustering in Graphs with Uncertain Structure

Joseph J. Pfeiffer, III  
Department of Computer Science  
Purdue University  
West Lafayette, IN 47907  
jpfeiffer@purdue.edu

Jennifer Neville  
Departments of Computer Science and Statistics  
Purdue University  
West Lafayette, IN 47907  
neville@cs.purdue.edu

Abstract

Much of the past work in network analysis has focused on analyzing discrete graphs, where binary edges represent the “presence” or “absence” of a relationship. Since traditional network measures (e.g., betweenness centrality) assume a discrete link structure, data about complex systems must be transformed to this representation before calculating network properties. In many domains where there may be uncertainty about the relationship structure, this transformation to a discrete representation will result in a lose of information. In order to represent and reason with network uncertainty, we move beyond the discrete graph framework and develop social network measures based on a probabilistic graph representation. More specifically, we develop measures of path length, betweenness centrality, and clustering coefficient—one set based on sampling and one based on probabilistic paths. We evaluate our methods on two real-world networks, Enron and Facebook, showing that our proposed methods more accurately capture salient effects without being susceptible to local noise.

Introduction

Much of the past work in network analysis has focused on analyzing discrete graphs, where entities are represented as nodes and binary edges represent the “presence” or “absence” of a relationship between entities. For example, network measures such as the average shortest path length and clustering coefficient have been used to explore the properties of biological and information networks (Watts and Strogatz 1998), while measures such as centrality have been used for determining the most important and/or influential people in social networks (Brandes 2001).

The main limitation of measures defined for a discrete representation is that they cannot easily be applied to represent and reason about uncertainty in the link structure. Link uncertainty may arise in domains where graphs evolve over time, as links observed at a earlier time may no longer be present or active at the time of analysis. In addition, there may be uncertainty with respect to the strength of the articulated relationships (Xiang, Neville, and Rogati 2010), or in other network domains (e.g., gene/protein networks) where relationships can only be indirectly observed. In this work, we formulate a probabilistic graph representation to analyze domains with these types of uncertainty.

The notion of probabilistic graphs have been studied previously. Notably, Frank (1969) has shown that for graphs with probability distributions over the weights for each edge, Monte Carlo methods can be used to sample to determine the shortest path probabilities between the edges. Then, Hua and Pei (2010) extends this to find the shortest weighted paths most likely to complete within a certain time constraint (e.g., the shortest distance across town in under half an hour). However, there has been little focus on how probabilistic paths and other graph structures should be incorporated into social network analysis measures.

Here, we develop analogs for three standard discrete graph measures—average shortest path length, betweenness centrality, and clustering coefficient—in the probabilistic setting. Specifically, we use probabilities on graph edges to represent link uncertainty and consider the distribution of possible (discrete) graphs that they define. Our first set of measures compute expected values over the distribution of graphs, sampling a set of discrete graphs from this distribution in order to efficiently approximate the path length, centrality, and clustering measures. We then develop a second set of measures that can be directly computed from the probabilities, which removes the need for graph sampling. This second approach focuses on the notion of the most probable paths in the network, rather than the shortest, and introduces a prior to incorporate the belief that the probability of successful information transfer is a function of path length.

We evaluate our measures on two real world networks: Enron email and Facebook micro communications, where the network transactions for each are associated with timestamps (e.g., email date). Thus we are able to compute the local and global measures at multiple time steps, where at each time step we consider the network information available up to and including t. We compare against two different approaches that use the discrete representation: an aggregate approach, which unions all previous transactions (up to t) into a discrete graph, and a slice approach, where only transactions from a small window (i.e., [t−d, t]) are included in the discrete representation. Our analysis shows that our proposed methods more accurately capture the salient changes in graph structure compared to the discrete methods without being susceptible to local, temporal noise.

Copyright © 2011, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
Sampling Probabilistic Graphs

Let $G = (V, E)$ be a graph where $V$ is a collection of nodes and $E \subseteq V \times V$ is the set of edges, or relationships, between the nodes. In order to represent and reason about relationship uncertainty, we associate each edge $e_{ij}$ (which connects node $v_i$ and $v_j$) with a probability $P(e_{ij})$. Then we can define $G$ to be a distribution of discrete, unweighted graphs. Assuming independence among edges, the probability of a graph $G \in \mathcal{G}$ is: $P(G) = \prod_{e_{ij} \in E} P(e_{ij}) \prod_{e_{ij} \notin E} [1 - P(e_{ij})]$. Note that although we assume edge independence for generation, this model can represent correlations in the graph structure by tying edge parameters (Leskovec et al. 2010). Since we have assumed edge independence, we can sample a graph $G_S$ from $G$ by sampling edges independently according to their probabilities $P(e_{ij})$. Based on this, we can develop methods to compute the expected shortest path lengths, betweenness centrality rankings, and clustering coefficients using sampling.

Calculating graph measures in this setting can be viewed generally as computing the expectation of a function $f$ over the distribution of graphs $G$. For any reasonable sized graph, the distribution $G$ will be intractable to enumerate explicitly, so to approximate the expected value of arbitrary functions we can sample from $G$. More specifically, we sample a graph $G_S$ by sampling edges uniformly at random according to their edge probabilities $P(e_{ij})$. Each graph that we sample in this manner has equal likelihood, thus we can draw $m$ sample graphs $G_S = \{G_1, ..., G_m\}$ and calculate the expected value for $f$ with the following: $E[G] = \sum_{G \in \mathcal{G}} f(G) \cdot P(G) \approx \frac{1}{m} \sum_{m} f(G_{m})$. $f$ can be any function over discrete, unweighted graphs.

In this paper we consider three social network measures: average shortest path length (SP), betweenness centrality (BC), and clustering coefficient (CC). Let $G \in \mathcal{G}$, $e_{ij} \in E$, $v_i, v_j \in V$ and $e_{ij} = \rho_{ij}$. Then we define the average shortest path length in $G$ as: $f_{SP}(G) = \frac{1}{|V|} \sum_{v_i \in V} \sum_{j \in V \neq v_i} \min_{e_{ij} \in E} |e_{ij}|$. Additionally, we define the betweenness centrality (BC) for a particular node $v_i$ as $f_{BC}(G) = \frac{1}{|V| \cdot (|V| - 1)} \sum_{v_i \in V} \sum_{v_k \in V \neq v_i \neq v_j} \sum_{e_{jk} \in E} P(e_{jk})$. The betweenness centrality ranking (BCR) for a node $v_i$ is then simply its index when all node BCs are ranked high to low. Lastly, we define the clustering coefficient (CC) for a node $v_i$ to be $f_{CC} = \frac{1}{|N_i|(|N_i| - 1)} \sum_{v_j \in N_i \setminus N_{v_i}} \sum_{v_k \in N_i \setminus N_{v_i} \setminus N_{v_j}} \Pr(e_{jk})$, where $N_i$ are the vertices $v_j$ such that $e_{ij} = 1$. More precise details are available in Pfeiffer and Neville (2011).

Probabilistic Path Length

In the previous section, we discussed an extension of discrete notions of shortest paths and centrality for a probabilistic graph framework, showing how to approximate expected values via sampling. However, since the expectation is over possible worlds (i.e., $G \in \mathcal{G}$), focusing on shortest paths may no longer be the best way to capture node importance. We note that previous work in the discrete framework (where all observed edges are equally likely) used shortest paths as a proxy for shortest importance. This implies a prior belief that shortest paths are more likely to be used successfully to transfer information and/or influence in the network. In domains with link uncertainty, the flow of information/influence will depend on both the existence of paths in the network and the use of those paths for communication/transmission. Thus, a measure that explicitly uses the edge probabilities to calculate most probable paths may more accurately highlight nodes that serve to connect many parts of the network. We discuss these issues more below.

Most Probable Paths To begin, we extend beyond discrete paths to consider probabilistic paths in our framework. Specifically, we calculate the probability of the existence of a path $\rho_{ij}$ as follows (again assuming edge independence): $P(\rho_{ij}) = \prod_{e_{uv} \in \rho_{ij}} P(e_{uv})$. Using path probabilities, we can now describe the notion of the most probable path. Given two nodes $v_i, v_j$, the most probable path is simply the one with maximum likelihood: $\rho_{ij}^{ML} = \arg\max \rho_{ij} P(\rho_{ij})$. We can compute the most likely paths in much the same way that shortest paths are computed on weighted discrete graphs, by applying Dijkstra’s shortest path algorithm, but instead of expanding on the shortest path, we expand the most probable path.

Transmission Prior Previous focus on shortest paths for assessing centrality relies on an implicit assumption that if an edge connects two nodes that it can be successfully used for transmission of information and/or influence in the network. Although prior work on information propagation in networks uses transmission probabilities, to our knowledge transmission probabilities have not previously been incorporated into node centrality measures. In our probabilistic framework, transmission probabilities can be incorporated to penalize the likelihood of longer paths in the graph. We conjecture that this approach will more accurately capture the role nodes play in the spread of information across multiple paths in the network.

To incorporate transmission likelihood into probabilistic paths, we assign a probability $\beta$ of success for every step in a particular path—corresponding to the probability that information is transmitted across an edge and is received by the neighboring node. If we denote $l$ to be the length of a path $\rho$, then we are interested in the case where all transmissions succeed, or $\beta^l$. Using this prior allows us to represent the expected probability of information spread in an intuitive manner, giving us a parameter $\beta$ which we can adjust to fit our expectations of information spread in the graph.

ML Handicapped Paths Combining the notion of probabilistic paths with an appropriate prior for modeling the probability of information spreading along the edges in the path, we can formulate the maximum likelihood handicapped path between two nodes $v_i$ and $v_j$ as follows: $\rho_{ij}^{MLH} = \arg\max \rho_{ij} \left[ P(\rho_{ij}) \cdot \beta^{l(\rho_{ij})} \right]$. To compute the most likely handicapped (MLH) paths, we follow the same formulation as the most probable paths, keeping track of the path length and posterior at each point. In the MLH formulation, probable paths are weighted by likelihood of transmission, thus nodes that lie on that are highly likely and relatively short, will have a high BC ranking. To calculate BCR ranking based on MLH paths, we can modify the Brandes
betweenness centrality algorithm (Brandes 2001), having it backtrack from the path that has the lowest probability of occurrence. Efficiency and the MLH relationship to discrete graphs can be found in Pfeiffer and Neville (2011).

**Probabilistic Clustering Coefficient**

We now outline a probabilistic measure of clustering coefficient that can be computed without sampling. If we again assume independent edges, the probability of triangle existence is equal to the product of the probabilities of the three sides. The expected number of triangles is then the sum of the triangles probabilities that include a given node \( v_i \). Denoting \( T_r \) to be the expected triangles including \( v_i \): 

\[
E_G[Tr] = \sum_{v_j,v_k \in N_i,v_j \neq v_k} [P(e_{ij}) \cdot P(e_{ki}) \cdot P(e_{jk})].
\]

Similarly we can denote \( Co_i \) to be the expected combinations (i.e., pairs) of the neighbors of \( v_i \) ans define the number of expected pairs as: 

\[
E_G[Co_i] = \sum_{v_j,v_k \in N_i,v_j \neq v_k} [P(e_{ij}) \cdot P(e_{ki})] \]

We can then define the probabilistic clustering coefficient to be the expectation of the ratio \( Tr_i/Co_i \), and approximate it via a first order Taylor expansion (Elandt-Johnson and Johnson 1980): 

\[
CC_v \approx E_G[Tr] / E_G[Co].
\]

Again, efficiency and relationships to discrete graphs can be found in Pfeiffer and Neville (2011).

**Experiments**

To investigate the performance of our proposed MLH and sampling methods for average path length, betweenness centrality and clustering coefficient, we compare to traditional baseline social network measures on data from Enron and Facebook. These datasets consist of time-stamped transactions among people (e.g., email, friend links). We will use the temporal activity information to derive probabilities for use in our methods, and evaluate our measures at multiple time steps to show the evolution of measures in the two datasets. For Enron, we consider the subset of the data comprised of the emails sent between employees, resulting in a dataset with 50,572 emails among 151 employees. The second dataset is from the Purdue University Facebook network. Specifically we consider one year’s worth of wall-to-wall postings between users in the class of 2011 subnetwork. The sample has 59,565 messages between 2,648 nodes—considerably larger than Enron.

We compare four network measures for each timestep \( t \) in each dataset. When evaluating at \( t \), each method is able to utilize the graph edges that have occurred up to and including \( t \). As baselines, we compare to (1) an aggregate method, which at a particular time \( t \) computes standard measures for discrete graphs (e.g., BCR) on the union of edges that have occurred up to and including \( t \), and (2) a time slice method, which again computes the standard measures, but only considers the set of edges that occur within the time window \( [t-\delta, t] \). For both Enron and Facebook we used \( \delta = 14 \) days.

We then compare to the sampling and MLH measures. For both the probabilistic methods, we use a measure of relationship strength based on exponentially decayed message counts as the edge probabilities for our analysis—note that any notion of uncertainty can be substituted at this step. We define the probability of an edge \( e_{ij} \) to be the likelihood that the two nodes \( v_i \) and \( v_j \) have an active relationship at the current timestep \( t_{now} \). The likelihood of activity is conditioned on having observed a communication message \( m_{ij} \) between the two nodes at time \( t(m_{ij}) \), where the impact of the message decays exponential in time: 

\[
P(e_{ij} | \lambda, m_{ij}) = \exp \left\{ -\frac{1}{\lambda} (t_{now} - t(m_{ij})) \right\}.\]

Assuming that we have \( k \) messages between \( v_i \) and \( v_j \), all of the messages \( m_{ij}, \ldots, m_{ij}^k \) contribute independently to relationship strength. Specifically, we define the probability of an active relationship to be \( 1 \) minus the probability that none of the observed messages indicate activity: 

\[
P(e_{ij} | \lambda, m_{ij}^{1:k}) = 1 - \prod_{t} (1 - P(e_{ij} | \lambda, m_{ij})).
\]

Figure 1: BCR of Lay and Skilling over time. Red lines indicate Skilling’s CEO announcement and resignation.

**Local Trend Analysis**

We analyze two key figures at Enron: Kenneth Lay and Jeffrey Skilling. These two were central to the Enron scandal—as first Lay, then Skilling, and then Lay again, assumed the position of CEO. The first event we consider (marked by a vertical red line in Fig. 1) is Dec. 13th, 2000, when it was announced that Skilling would assume the CEO position at Enron, with Lay retiring but remaining as a chairman (Marks ). In Figure 1, both the sampling method and the MLH method identify a spike in BCR for Lay and Skilling directly before the announcement. This is not surprising, as presumably Skilling and Lay were informing the other executives about the transition that was about to be announced. Following the transition, both probabilistic methods agree that Skilling and Lay have lower centrality. The time slice method (1.c) produces no change in Lay’s BCR, despite his central role in the transition. Also, there are a few random spikes in Skilling’s BCR, which illustrates the variance that results from using the time slices. The aggregate model (1.d) fails to reduce Skilling’s BCR to the expected levels following the announcement—although it is still fairly early in the time window, the aggregate method is unable to track current events based on its unioning of all past transactions.

The second event we consider (marked by the 2nd ver-
Global Trend Analysis

In this paper we investigated the problem of calculating centrality and clustering in networks with edge uncertainty. We introduced sampling-based measures for average shortest path and betweenness centrality, as well as measures based on most probable paths, which are more intuitive for capturing network flow. We outlined exact methods to compute most probable paths (and by extension, most probable betweenness centrality), and incorporated a transmission probability to capture the notion of influence across uncertain paths. In addition, we outlined a probabilistic version of clustering coefficient and gave a first order Taylor expansion approximation for computation. We analyzed our proposed methods using time evolving networks from Enron and Facebook. We demonstrated the limitations of using either an aggregate graph representation or a slice-based representation in networks with uncertainty due to evolution over time, namely that the aggregate approach fails to react to changes in network structure and that the slice approach exhibits extreme variability due to temporal noise. The results provide empirical evidence to illustrate the utility of the probabilistic sampling and MLH-based social network measures. In particular, the centrality rankings for the Enron employees match our intuitions based on knowledge of the Enron timeline.

Acknowledgements

This material is based in part upon work supported by IARPA under AFRL contract number FA8650-10-C-7060. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL or the U.S. Government. Pfeiffer is supported by a Purdue Frederick N. Andrews Fellowship.

References

Brandes, U. 2001. A faster algorithm for betweenness centrality. Journal of Mathematical Sociology 25:163–177.
Elundt-Johnson, R. C., and Johnson, N. L. 1980. Survival models and data analysis. John Wiley & Sons, New York.
Frank, H. 1969. Shortest paths in probabilistic graphs. In Operations Research, 17, No. 4 (Jul. - Aug., 1969), pp. 583-599.
Hua, M., and Pei, J. 2010. Probabilistic path queries in road networks: traffic uncertainty aware path selection. In EDBT, 347–358.
Leskovec, J.; Chakrabarti, D.; Kleinberg, J.; Faloutsos, C.; and Ghahramani, Z. 2010. Kronecker graphs: An approach to modeling networks, JMLR.
Leskovec, J.; Kleinberg, J.; and Faloutsos, C. 2005. Graphs over time: Densification laws, shrinking diameters and possible explanations. In KDD.
Marks, R. Enron timeline. http://www.agsm.edu.au/bobm/teaching/BE/Enron/timeline.html.
Pfeiffer III, J. J., and Neville, J. 2011. Methods to determine node centrality and clustering in graphs with uncertain structure. Technical Report 11-010, CS Dept, Purdue University.
Watts, D. J., and Strogatz, S. H. 1998. Collective dynamics of ‘small-world’ networks. Nature 393(6684):440–442.
Xiang, R.; Neville, J.; and Rogati, M. 2010. Modeling relationship strength in online social networks. In WWW 2010.