2011

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

Joseph J. Pfeiffer III
Purdue University, jpfeiffer@purdue.edu

Jennifer Neville
Purdue University, neville@cs.purdue.edu

Report Number:
11-010

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 47909
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) utilize a discrete link structure, complex systems must be transformed to this representation in order to investigate network properties. However, in many domains there may be *uncertainty* about the relationship structure and any uncertainty information would be lost in translation to a discrete representation. Uncertainty may arise in domains where there is moderating link information that cannot be easily observed, i.e., links become inactive over time but may not be dropped or observed links may not always corresponds to a valid relationship. In order to represent and reason with these types of 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 three real-world networks from Enron, Facebook, and DBLP, showing that our proposed methods more accurately capture salient effects without being susceptible to local noise, and that the resulting analysis produces a better understanding of the graph structure and the uncertainty resulting from its change over time.

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. Complex systems of relationships are first transformed to a discrete graph representation (e.g., a friendship graph) and then the connectivity properties of these graphs are used to investigate and understand the characteristics of the system. 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; Leskovec, Kleinberg, and Faloutsos 2005), while measures such as centrality have been used for determining the most important and/or influential people in social networks (Freeman 1977; Bonacich 1987).

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 the time of analysis. For example in online social networks, users articulate "friendships" with other users and these links often persist over time, regardless of whether the friendship is maintained. This can result in uncertainty about whether an observed friendship link is still *active* at some later point in time. In addition, there may be uncertainty with respect to the *strength* of the articulated relationships (Xiang, Neville, and Rogati 2010), which can result in uncertainty about whether an observed relationship will be used to transmit information and/or influence. Furthermore, there are other network domains (e.g., gene/protein networks) where relationships can only be indirectly observed so there is uncertainty about whether an observed edge (e.g., protein interaction) actually indicates the presence of a valid relationship.

In this work, we formulate a probabilistic graph representation to analyze domains with these types of uncertainty and develop analogues 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, then we develop measures that consider the properties of the graph population defined by this distribution.

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. The second approach also affords us the opportunity to consider more than just shortest paths in the network. We note that previous focus on shortest paths is due in part to an implicit belief that short paths are more likely to result in successful transfer of information and/or influence between two nodes. This has led other works to generalize shortest paths to the probabilistic domain for their own purposes (Potamias et al. 2009). However, in a probabilistic

framework we can also directly compute the likelihood of a path and consider the *most probable* paths, which are likely to facilitate information flow in the network.

With probabilistic paths, we also introduce a *prior* to incorporate the belief that the probability of successful information transfer is a function of path length—since the existence of a relationship does not necessarily mean that information/influence will be passed across the edge. This formulation, which models the likelihood of information spread throughout the graph, is consistent with the finding in (Onnela et al. 2007), which identified that constricting and relaxing the flow along the edges in the network was necessary to model the true patterns of information diffusion in an evolving communication graph.

We evaluate our measures on three real world networks: Enron email, Facebook micro communications, and DBLP coauthorships. In these datasets, the network transactions are each associated with timestamps (e.g., email date). Thus we are able to compute the local (node-level) and aggregate (graph-level) measures at multiple time steps, where at each time step t 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., $[t-\delta, t]$) are included in the discrete representation. For our methods, we estimate edge probabilities from the transactions observed up to t, weighting each transaction with an exponential decay function. 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. Thus the resulting analysis produces a better understanding of the graph structure and its change over time.

Related Work

The notion of probabilistic graphs have been studied previously, notably by (Frank 1969), (Hua and Pei 2010) and (Potamias et al. 2009). (Frank 1969) showed how 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. (Hua and Pei 2010) then 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). In (Potamias et al. 2009), the most probable shortest paths are used to estimate the k-nearest neighbors in the graph for a particular node. Although (Potamias et al. 2009) draws sample graphs based on *likelihood* (i.e., sampling each edge according to its probability), in their estimate of the shortest path distribution they weight each sample graph based on its probability, which is incorrect unless the samples are drawn uniformly at random from the distribution. In this work, we sample in the same manner as (Potamias et al. 2009), but weight each sample uniformly in our expectation calculations—since, when the graphs are drawn from the distribution based on their likelihood, the graphs with higher likelihood are more likely to be sampled.

There has also been some recent work that has developed measures for time-evolving graphs, e.g., to identify the most central nodes throughout time (Tang et al. 2010) and identify the edges that maximize communication over time (Kossinets, Kleinberg, and Watts 2008). However, these works fail to account for the uncertainty in both the link structure and the the communication across links (as users are unlikely to propagate all information across a single edge). Our use of a probabilistic graph framework and transmission prior address these two cases of uncertainty.

Sampling Probabilistic Graphs

Let $G = \langle V, E \rangle$, be a graph where V is a collection of nodes and $E \in V \times V$ 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 $\mathcal{G}$ to be a distribution of discrete, unweighted graphs. Assuming independence among edges, the probability of a graph $G \in \mathcal{G}$ is: $P(G) = \prod_{e_{ij} \in E} P(e_{ij}) \prod_{e_{ij} \notin E} [1 - P(e_{ij})]$. Since we have assumed edge independence, we can sample a graph G_S from $\mathcal{G}$ 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.

Probabilistic Average Shortest Path Length Let $\rho_{ij} = \{v_{k_1}, v_{k_2}, ..., v_{k_q}\}$ refer to a *path* of q vertices connecting two vertices v_i and v_j, i.e., $v_{k_1} = v_i$ and $v_{k_q} = v_j$, and from each vertex to the next there exists an edge: $e_{k_i k_{i+1}} \in E$ for $i = [1, q-1]$. Let $V(\rho_{ij})$ and $E(\rho_{ij})$ refer to the set of vertices and edges respectively, in the path and let $|\rho_{ij}| = |E(\rho_{ij})|$ refer to the *length* of the path. Assuming connected graphs, for every unweighted graph $G = \langle V, E \rangle \in \mathcal{G}$ there exists a shortest path ρ_{ij}^{min} between every pair of nodes $v_i, v_j \in V$. Letting $\text{SP}_{ij} = |\rho_{ij}^{min}|$, we can then define the average shortest path length in G as: $\overline{\text{SP}}(G) = \frac{1}{|V| \cdot (|V|-1)} \sum_{i \in V} \sum_{j \in V; j \neq i} \text{SP}_{ij}$.

Now, when there is uncertainty about the edges in G, we can compute the *expected* average shortest path length by considering the distribution of graphs $\mathcal{G}$. For any reasonable sized graph, the distribution $\mathcal{G}$ will be intractable to enumerate explicitly, so instead we sample from $\mathcal{G}$ to approximate the expected value. 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 shortest path length with the following:

$$\mathbb{E}_{\mathcal{G}}\left[\,\overline{\text{SP}}\,\right] = \sum_{G \in \mathcal{G}} \overline{\text{SP}}(G) \cdot P(G) \simeq \frac{1}{m} \sum_{m} \overline{\text{SP}}(G_m) \quad (1)$$

Since the sampled graphs are unweighted, it takes $O\left(|V|\,|E|\right)$ time to compute $\overline{\text{SP}}$ for each sample (Brandes 2001). This results in an overall cost of $O\left(m \cdot |V|\,|E|\right)$ to compute $\mathbb{E}_{\mathcal{G}}\left[\,\overline{\text{SP}}\,\right]$.

Sampled Centrality Betweenness centrality for a node v_i is defined to be the number of shortest paths between other pairs of nodes which pass through v_i: $BC_i = |\{\rho_{jk}^{min} \in G : v_i \in V(\rho_{jk}) \ \wedge \ i \neq j,k\}|$. Vertices that contribute to the existence of many shortest paths will have a higher BC score than other nodes that contribute to fewer shortest paths, thus BC is used a measure of importance or centrality in the network. It is difficult to directly compare BC values across graphs since the number of shortest paths varies with graph size and connectivity. Thus, typically analysis focuses on *betweenness centrality rankings* (BCR), where the nodes are ranked in descending order of their BC scores and the node with the highest BC score is given a BCR of 1.

As discussed above, we can compute the shortest paths for each unweighted graph $G \in \mathcal{G}$, then we can also compute the BCR values for each unweighted graph $G \in \mathcal{G}$. We denote $\text{BCR}_i(G)$ as the betweenness centrality ranking for node v_i in G. Then we can approximate the expected BCR for each node by sampling a set of m graphs from $\mathcal{G}$:

$$\mathbb{E}_{\mathcal{G}}[\text{BCR}_i] \simeq \frac{1}{m} \sum_m \text{BCR}_i(G_m) \tag{2}$$

Again, since the sampled graphs are unweighted, it takes $O(|V||E|)$ time to compute the BCR for each sample (Brandes 2001), resulting in an overall cost of $O(m \cdot |V||E|)$.

Sampled Clustering Coefficients Clustering coefficient is a measure of how the nodes in a graph cluster together (Watts and Strogatz 1998). For a node v_i with $N_i = \{v_{j_1}, ..., v_{j_n}\}$ neighbors (e.g., $e_{ij_1} \in E$), its clustering coefficient is defined as $\text{CC}_i = \frac{1}{|N_i|(|N_i|-1)} \sum_{v_j \in N_i} \sum_{v_k \in N_i, k \neq j} \mathbb{I}_E(e_{jk})$, where $\mathbb{I}_E$ is an indicator function which returns 1 if v_j is connected to v_k. CC can be thought of as the fraction connected pairs of neighbors of v_i. We denote $\text{CC}_i(G)$ as the clustering coefficient for node v_i in graph G. Similar to paths, we can compute clustering coefficients for every graph $G \in \mathcal{G}$. Thus we can approximate the expected CC for each node by sampling a set of m graphs from $\mathcal{G}$:

$$\mathbb{E}_{\mathcal{G}}[\text{CC}_i] \simeq \frac{1}{m} \sum_m \text{CC}_i(G_m) \tag{3}$$

Under the assumption that the maximum degree in the graph can be bounded by a fixed constant (which is typical for sparse social networks), we can compute the clustering coefficient for a single graph in $O(|V|)$ time (i.e., $O(1)$ for each node), which results in an overall cost of $O(m \cdot |V|)$.

Probabilistic Path Length

In the previous section, we discussed how to extend the discrete notions of shortest paths and centrality into a probabilistic graph framework via expected values, and we showed how to estimate approximate values using sampling. While our sampling-based measures are valid and give informative results (see section 6 for details), they have two limitations which restrict their applicability.

First, the effectiveness of the approximation depends on the number of samples from $\mathcal{G}$. We note that (Potamias et al. 2009) used a Hoeffding Inequality to show that relatively few samples are needed to compute an accurate estimate of independent shortest paths in probabilistic graphs. However, since our the calculation of BCR is based on the joint occurrence of shortest paths in the graph, this bound will not hold for our measures.

Second, since the expectation is over possible worlds (i.e., $G \in \mathcal{G}$), the focus on shortest paths may no longer be the best way to capture node *importance*. We note that in the discrete framework, where all edges are equally likely, the use of shortest paths as a proxy for importance implies a prior belief that shorter 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. In a probabilistic framework, we have an opportunity to explicitly incorporate the latter, by encoding our prior beliefs about transmission likelihood into measures of node importance. Furthermore, although a probabilistic representation enables analysis of more than just shortest paths, as we note above, even to capture shortest paths the sampling methods described previously may need many samples to accurately estimate the joint existence of shortest paths. 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 each of these issues more below.

Most Probable Paths To begin, we extend the notion of discrete paths to probabilistic paths in our framework. Specifically, we can calculate the probability of the existence of a path ρ_{ij} as follows (again assuming edge independence): $P(\rho_{ij}) = \prod_{e_{uv} \in E(\rho_{ij})} P(e_{uv})$. Using the path probabilities, we can now describe the notion of the *most probable* path. Given two nodes v_i, v_j, the most probable path path is simply the one with *maximum likelihood*: $\rho_{ij}^{ML} = \text{argmax}\, P(\rho_{ij})$. 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. Thus, all most probable paths can be calculated in $O\left(|V||E| + |V|^2 \log |V|\right)$.

Transmission Prior Previous focus on shortest paths for assessing centrality points to 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 there has been work both in maximizing the spread of information in a network through the use of central nodes (Boragatti 2005; Newman 2005) and in the study of information propagation through the use of transmission probabilities (Goldenberg, Libai, and Muller 2001), there has been little prior work that has incorporated transmission probabilities into node centrality measures. Centrality measures based on random walks and eigenvectors (Newman 2005) implicitly penalize longer paths as they consider *all*

paths between nodes in the network. However, in our framework we can incorporate transmission probabilities to penalize the probabilities of longer paths in the graph, in order to more accurately capture the role nodes play in the spread of information across multiple paths in the network.

Consider the case where there is one path of nine people where each edge has high probability of existence (e.g., 0.95) and another path of three people where the edge probabilities are all moderate (e.g., 0.70), both ending at node v. Here, the longer path is more likely to exist than the shorter path, but in this example we are more interested in which path is used to transfer a virus to v. Even when an edge exists (i.e., the relationship is active), the virus will not be passed with certainty to the next node, thus the *transmission probability* is independent of the edge probability. Moreover, when the transmission probability is less than 1, it is more likely that the virus will be transmitted across the shorter path, since the longer path presents more opportunities for the virus to be dropped. This provides additional insight as to why shortest paths have always been considered important—there is generally a higher likelihood of transmission if it is passed through fewer nodes in the network.

To incorporate transmission likelihood into our probabilistic paths, we assign a probability β 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 ρ, and s to be the number of successful transmissions along the path, we can use a binomial distribution to represent the transmission probability across ρ with:

$$\text{SBin}(s|\beta) = \text{Bin}(s = l|l, \beta) = \beta^l$$

Here SBin corresponds to the case where the transmission *always* succeeds (i.e., across all edges in ρ). Using this binomial distribution as a prior allows us to represent the expected probability of information spread in an intuitive manner, giving us a parameter β which we can adjust to fit our expectations for the information spread in the graph. Note that setting $\beta = 1$ is equivalent to the most probable paths discussed earlier. The prior effectively *handicaps* longer paths through the graph. Although, there is a correlation between shortest (certain) paths and handicapped (uncertain) paths, these formulations are *not* equivalent, since the latter produces a different set of paths when the shortest paths have low probability of existence.

ML Handicapped Paths Now that we have both the notion of a probabilistic path, and 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 to be:

$$\rho_{ij}^{MLH} = \underset{\rho_{ij}}{\text{argmax}} \left[P(\rho_{ij}) \cdot \text{SBin}(\,|\rho_{ij}|\,|\,\beta)\right] \qquad (4)$$

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 paths that are highly likely and relatively short, will have a high BC ranking. To calculate BCR ranking based on MLH paths, we can use a weighted betweenness centrality algorithm. Specifically, we modify Brandes' algorithm (Brandes 2001) to start with the path that has the lowest probability of occurrence to be the one to backtrack from, enabling computation of the betweenness centrality in $O\left(|V|\,|E| + |V|^2 \log |V|\right)$.

Comparison with Discrete Graphs

The formulation of MLH Paths has inherent benefits, most notably with its direct connection to the previously well-studied notions of shortest paths and betweenness centrality in discrete graphs. In fact, we can view a discrete graph G as being a special case of probabilistic graph with edge probabilities:

$$P(e_{ij}) = \begin{cases} 1 & \text{if an edge exists} \\ 0 & \text{if the edge does not exist} \end{cases} \qquad (5)$$

We denote the distribution of graphs defined by these probabilities as $\mathcal{G}_1$. Note that the only graph in $\mathcal{G}_1$ with non-zero probability is G—since if an edge exists in a discrete graph, then it exists with complete certainty, likewise, if an edge is not present, we are certain it does not exist, thus $P(G) = 1$.

Theorem 1. *For every pair of nodes v_i and v_j, the shortest path in the discrete graph ($\rho_{ij} \in G$) is equal to the most probable path discovered by the MLH algorithm ($\rho_{ij}^{MLH} \in \mathcal{G}_1$), for $0 < \beta < 1$.*

Proof. In $\mathcal{G}_1$ every $P(e_{ij})$ is either 1 or 0, thus every case where $P(\rho_{ij}) > 0$ is precisely $P(\rho_{ij}) = 1$. If we choose the shortest path from the discrete graph, it will have length $l^* = |\rho_{ij}|$, and the MLH probability for the same path will be β^{l^*}. Clearly, if a longer path were chosen by MLH, its probability would be less than β^{l^*}, and we know that no shorter paths exist—since all paths shorter than ρ_{ij} would involve an edge than did not exist in G and thus would have probability 0 . □

Corollary 1. *The betweenness centrality using shortest paths on a discrete graph G can be equivalently calculated with most probable handicapped paths over $\mathcal{G}_1$, where edge probabilities are defined by Equation 5.*

Proof. This follows directly from Thm 1. □

Probabilistic Clustering Coefficient

We now outline a probabilistic measure of clustering coefficient that can be computed without the need for sampling. If we assume independence between edges, the probability of a triangle's 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 Tr_i to be the expected triangles including v_i: $\mathbb{E}_{\mathcal{G}}\left[\text{Tr}_i\right] = \sum_{v_j, v_k \in N_i, v_j \neq v_k} [P\left(e_{ij}\right) \cdot P\left(e_{ki}\right) \cdot P\left(e_{jk}\right)]$. Denoting Co_i to be the expected *combinations* (i.e., coexisting pairs) of the neighbors of v_i, we then get: $\mathbb{E}_{\mathcal{G}}\left[\text{Co}_i\right] = \sum_{v_j, v_k \in N_i, v_j \neq v_k} [P\left(e_{ij}\right) \cdot P\left(e_{ki}\right)]$. 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):

$$\text{CC}_i = \mathbb{E}_{\mathcal{G}}\left[\frac{\text{Tr}_i}{\text{Co}_i}\right] \approx \frac{\mathbb{E}_{\mathcal{G}}\left[\text{Tr}_i\right]}{\mathbb{E}_{\mathcal{G}}\left[\text{Co}_i\right]} \tag{6}$$

Assuming again that the maximum degree in the graph can be bounded by a fixed constant, we can compute the probabilistic clustering coefficient in $O(|V|)$ time ($O(1)$ for each node). Additionally, the probabilistic approximation to the clustering coefficient shares connections with the traditional clustering coefficients on discrete graphs.

Theorem 2. *The probabilistic clustering coefficients computed in $\mathcal{G}_1$, with probabilities defined by 5 for a discrete graph G, are equal to the discrete clustering coefficients calculated on G.*

Proof. Any triangle from G has probability 1 in $\mathcal{G}_1$, while any non-triangle in G clearly has probability 0. The same is true for the combinations of pairs of neighbors. As such, the sums of the numerators and denominators will be equal for both clustering coefficient. □

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, DBLP, and Facebook. These datasets all consist of time-stamped *transactions* among people (e.g., email, joint authorship). 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 three datasets.

Datasets

For our analysis we first use the Enron dataset (Shetty and Adibi 2004). The advantage to this dataset is that it allows us to understand the effects of our probabilistic measures because key events and central people have been well documented (Marks). 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.

Our second dataset is a sample from the DBLP computer science citation database. We considered the set of authors who had published more than 75 papers in the timeframe 1967-2006, and the coauthor relationships between them. The resulting subset of data consisted of 1,384 nodes, with 23,748 co-authors relationships.

Our third 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 2,648 nodes with 59,565 messages.

Methodology

We compare four network measures for each timestep t in each dataset. When evaluating at time 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 the Enron and Facebook, we used $\delta = 14$ days and for DBLP, we considered $\delta = 1$ year.

We then compare to the sampling and MLH measures. For both the probabilistic methods, we need a measure of relationship strength to use as probabilities in our model. Although any notion of relationship strength can be substituted at this step, in this work we utilize a measure of relationship strength based on decayed message counts. More specifically, we define two separate and distinct notions of connection between nodes: *edges* and *messages*. We define an edge e_{ij} to be the unobservable probabilistic connection between two nodes, indicating whether the nodes have an active relationship. This is in contrast to messages: a message m_{ij} is a concrete and directly measurable communication between two nodes v_i and v_j, such as a wall posting or email, occurring at a specific time, which we denote $t(m_{ij})$. We define the probability of of nodes v_i and v_j having an *active* relationship at the current timestep t_{now}, based on observing a message at time $t(m_{ij})$, to be the exponential decay of a particular message:

$$P\left(e_{ij}^t | m_{ij}\right) = \text{Exp}\left(m_{ij} | t_{now}, \lambda\right) = \exp\left\{-\frac{1}{\lambda}\left(t_{now} - t\left(m_{ij}\right)\right)\right\}$$

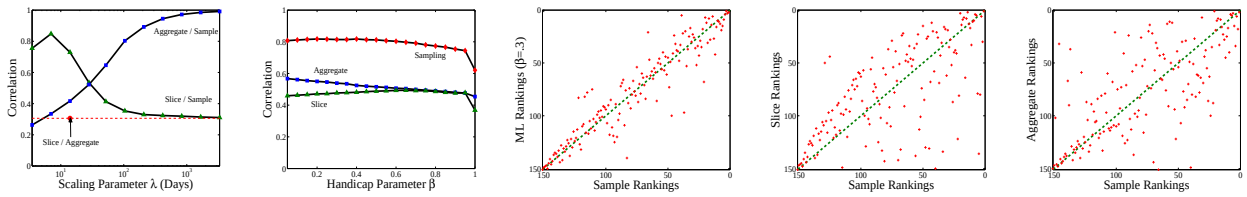
Note that the *scaling* parameter λ refers to the adjustment of the basic time unit (e.g. 7 days to 1 week), not the *rate* parameter which defines the exponential probability density function, which in this case is 1. This allows for assigning a probability of 1 to the case when $t\left(m_{ij}\right) = t_{now}$, but it also assigns reasonable probabilities (i.e., slows the decay) for messages that happened in the recent past, which could still indicate active relationships.

Now, we assume we have k messages between v_i and v_j, and any of the messages $m_{ij}^1, \ldots, m_{ij}^k$ can contribute to the relationship strength, which is defined to be 1 minus the probability that none of them contribute:

$$P\left(e_{ij}^t | m_{ij}^1, \ldots, m_{ij}^k\right) \;=\; 1 - \prod_k \left(1 - \text{Exp}\left(m_{ij}^k | t_{now}\right)\right)$$

In order to choose a scaling parameter λ for the exponential decay, we measured the average correlation from the sampling method BCR against the time slice ranking and aggregate method for each Enron employee, for different values of λ (see Figure 1.a). Note that a λ close to 0 corresponds to 'forgetting' a transaction quickly and is highly correlated with the slice method, while a large λ corresponds to 'remembering' a transaction for a long time, giving it high correlation with the aggregate method. In order to balance between short term change and long term trends we set λ to a 'middle ground' with $\lambda = 28$ days. This applies to both the Enron and Facebook datasets. For DBLP, where we evaluate yearly, λ is set to 2 years to keep the ratio between time slice and λ consistent between Facebook, Enron, and DBLP.

In order to choose a value for the β parameter in the MLH method, we measured the average correlation of the BCR from the MLH method and compared them to the sampling, aggregate, and slice rankings for different values of β. We



(a) (b) (c) (d) (e)

Figure 1: (a) Correlation between methods for varying values of λ. (b) Correlation of MLH with other methods as β is varied. (c-e) Correlations of Enron employee BCRs across methods, for the time segment ending August 24^{th}, 2001

can see in Figure 1.b that as long as β is non-zero, it has minimal effect on the correlations. For the experiments reported in this paper, we set $\beta = .3$. Note that omission of the prior (i.e., $\beta = 1$) in will make the MLH paths similar to the slice paths, with added paths between vertices which are disjoint in a particular time slice.

The final parameter setting is the number of samples to consider in each of sampling-based measures. Earlier we discussed how we are computing the joint instances of shortest paths, and that the bound by (Potamias et al. 2009) does not hold. Due of this, we exploit the small size of the Enron dataset and take 10,000 samples; however, with the two larger graphs we use a smaller sample size of 200 in order to make the experiments tractable.

Method Correlations on Enron Data

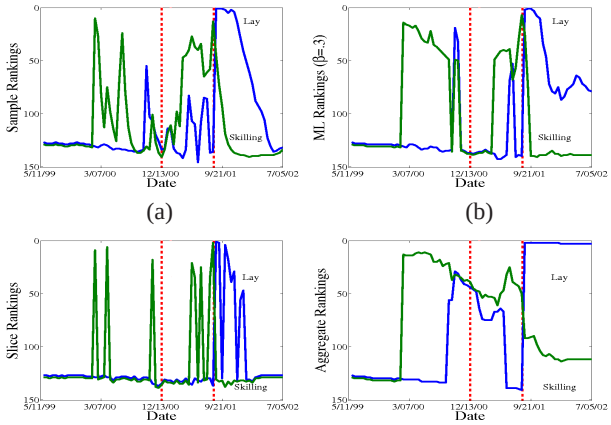
In order to illustrate the differences between the four methods, we analyze their respective BCR on the Enron data for the time window ending August 14^{th}, 2001. Figure 1.c-e shows the correlations of employee BCR across a pair of methods: points on the diagonal green line indicate 'perfect' correlation between the rankings of two methods.

Figure 1.c shows that the MLH method closely matches the sampling method, with only a few nodes varying from the diagonal. However, a large number of nodes that the sampling method determines to have high centrality are missed by the slice method, due to the slice's inability to see transactions that occurred prior to the evaluation time window. Additionally, we note that August 14^{th}, 2001 is relatively late in the Enron timeline, which results in the aggregate method having little correlation with the sampling method, since the more recent changes are washed out by past transactions in the aggregate approach.

Local Trend Analysis

Lay and Skilling Here, we analyze two key figures at Enron: Kenneth Lay and Jeffery Skilling. These two were central to the Enron scandal—as first Lay, then Skilling, and then Lay again, assumed the position of CEO. We can analyze the BCR for Lay and Skilling during these transition periods, as we expect large changes to affect both of them.

The first event we consider (marked by a vertical red line in Figure 2) is *December 13^{th} 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 2.a, both the sampling method and the MLH method identify a spike in BCR for both Lay and Skilling directly



(a) (b) (c) (d)

Figure 2: BCR of Lay and Skilling over time. Red lines indicate Skilling's CEO announcement and resignation.

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.

The time slice method (2.c) produces no change in Lay's BCR, despite his central role in the transition. Skilling shows a few random spikes of BCR, which illustrates the variance associated with using the time slices. The aggregate model (2.d) fails to reduce Skilling's BCR to the expected levels following the announcement—this is fairly early in time and we are already seeing the aggregate method's inability to track current events based on its union of all past transactions. Both the sampling method and the MLH methods capture this; MLH has him return to an extremely low centrality, while sampling has fairly low with some variance.

The second event we consider (marked by the 2nd vertical red line in Figure 2) is *August 14^{th} 2001*, when, seven months after initially taking the CEO position, Skilling approached Lay about resigning (Marks). During the entirety of Skilling's tenure, we see that Lay has a slight effect on the sample rankings but is not what would be considered a 'central' node. Not surprisingly, Skilling has a fairly high centrality during his time as CEO; both the sampling method and MLH method capture this.

Prior to the announcement of Lay's takeover as CEO, the slice method still had no weight on him, despite his previous involvement with the first transition. Also, we note that the sampling, MLH, and slice methods all agree that after Lay's initial spike from the Skilling resignation, he resumes having a lower centrality, which the aggregate method misses.

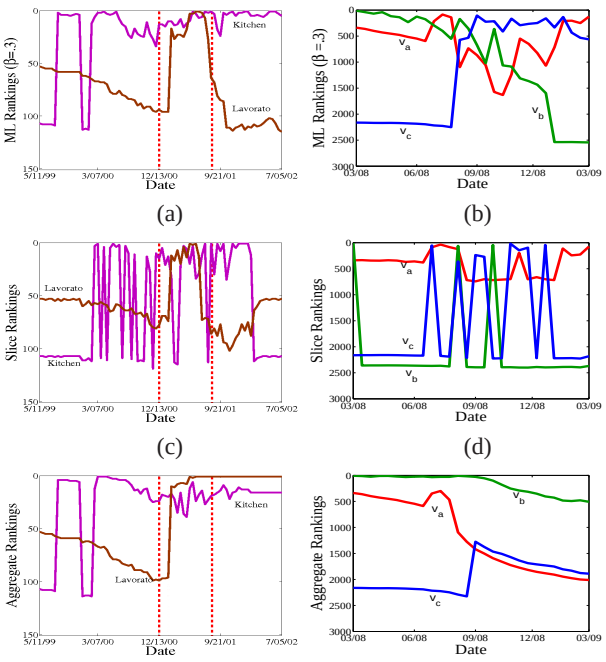


Figure 3: (a,c,e) BCR of Kitchen and Lavorato. (b,d,e) BCR for 3 nodes in the Purdue Facebook network.

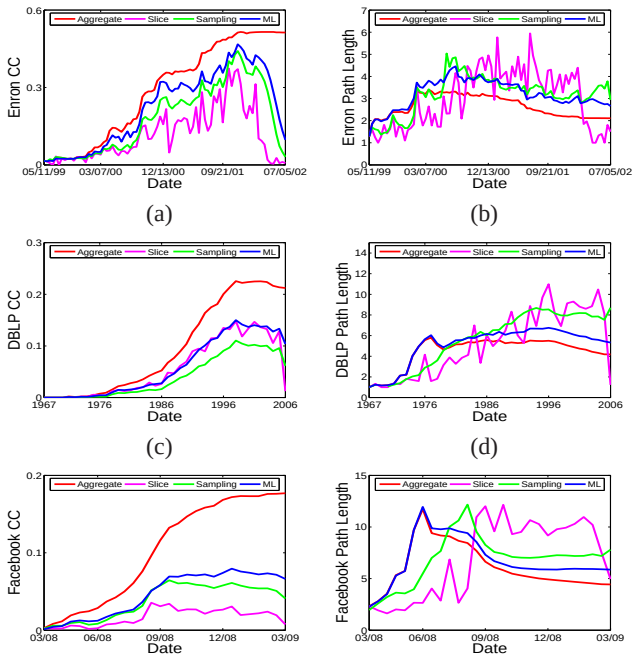


Figure 4: Average path lengths and clustering coefficients for Enron (a,b), DBLP (c,d) and Facebook (e,f).

In general, the sampling method seems to mirror the slice method, albeit with less variance, but it not as smooth as the MLH method, indicating the utility of considering most probable paths.

Kitchen and Lavorato Next we analyze Louise Kitchen and John Lavorato, who were executives (Shetty and Adibi 2004) for Enron Americas, which was the wholesale trading section of Enron (Raghavan, Kranhold, and Barrionuevo). They are notable because of the extraordinarily high bonuses they received as Enron was being investigated, and were also found to have a high temporal betweenness centrality using the method defined by (Tang et al. 2010). We can see in Figure 3 (a,c,e) the rankings of Kitchen and Lavorato, and can see the benefit of using the probabilistic framework's ability to key in on centralities at *specific* times, rather than using the temporal definition *through* time proposed by (Tang et al. 2010). We see that while Lavorato might have gotten a large bonus, he is *only* important during Skilling's tenure as CEO; his centrality drops noticeably otherwise. On the other hand, Kitchen had extremely high rankings throughout.

Here, we see that the slice method exhibits high variability, especially with Kitchen, while the aggregate cannot recognize Lavorato's lack of importance after Skilling's departure. The MLH method is able to smoothly capture Kitchen's centrality, while keeping Lavorato important solely during Skilling's CEO tenure.

Facebook Centrality Unlike the Enron dataset, the Purdue Facebook dataset does not have well-established ground truths, where we can use the known characteristics and behaviors of particular nodes for evaluation. However, we can examine aspects of a few representative nodes to illustrate the problems that lie with usage of the aggregate or static methods. First, we can see from Figure 3.d that v_a (red) has a consistently high ranking in the slice method, which the MLH method captures (3b). However, this person has a declining ranking in the aggregate method, as the aggregate is unable to capture current events—past information in the aggregate graph results in many paths that bypass v_a, missing this central node in later timesteps.

The next person we consider is denoted by v_b (green). In 3.d, we can see that the slice method initially identifies this person as having high centrality, then their BCR bottoms out, and then peaks a few times again approximately midway through the timeline. The MLH method also initially identify v_b as central, with a degradation over time. In contrast, the aggregate method fails to detect the inactivity later in the timeframe and continues to give v_b a high centrality ranking throughout the entire time window.

The final person we consider is denoted by v_c (blue) in Figure 3. We can see in 3.d that the slice method exhibits large variability for v_c, but that there are many slices in the middle to end of the timeframe where the node is identified as highly central. The aggregate method is unaware of this activity and ranks v_c at a relatively low level throughout the timeseries. In contrast, the MLH method is able to recognize the node's growing importance as time evolves, and do so much more smoothly than the slice method (3.d). In doing so, the MLH method can find instances of high centrality when both discrete methods fail.

Global Trend Analysis

In Figure 4, we report the average path lengths for the various measures: MLH paths, probabilistic shortest paths, the

aggregate shortest paths and the slice shortest paths. Additionally, we report the average sampled clustering coefficient, the clustering coefficient approximation, and the aggregate and slice discrete clustering coefficients. These are done for each of the three datasets through time, and we investigate changes in these global statistics to understand what, if any, changes occur with respect to the *small world* network structure of the data (Watts and Strogatz 1998).

In Figures 4.a,c,e, we show the clustering coefficients for each of the three datasets. The aggregate graph significantly overestimates the amount of *current* clustering in the graph, while the slice method is highly variable, especially for Enron. In general, both probabilistic measures are in between the two extremes, balancing the effects of recent data and decreasing the long term effect of past information, with the MLH performing similarly to the sampled clustering coefficient, and even better on DBLP, where sampling undercuts the clustering (likely due to small sample size).

Next, in Figures 4.b,d,f, we examine the *shrinking diameter* of these small world networks (Leskovec, Kleinberg, and Faloutsos 2005). Here, the aggregate underestimates the path length at a current point in time. We can see that the most probable paths closely follows the sampling results, with both lying between the slice and aggregate measures while avoiding the variability of the slice method.

Conclusions

In this paper we investigated the problem of calculating centrality and clustering in an uncertain network, and analyzed our methods using time evolving networks. We demonstrated the limitation of using an aggregate graph representation to capture uncertainty in the network structure due to changes over time, as well as the limitation of using a slice-based representation due to its extreme variability. We introduced sampling-based measures for average shortest path and betweenness centrality, as well as measures based on the most probable paths, which are more intuitive for capturing network flow. We also outlined exact methods for the computation of most probable paths (and by extention, most probable betweenness centrality), and incorporated the notion of transmission probability. Additionally, we developed a probabilistic clustering coefficient and gave a first order Taylor expansion approximation for computation.

We provided empirical evidence on the Enron, DBLP, and Facebook datasets showing the sampling and MLH's intuitive centrality rankings for the Enron employees and Facebook members, as well as the global properties for all three. The probabilistic centrality and clustering formulations are inherently smoother than the measures computed from discretized time slices, however they can reason about *likely* change in graph structure due to changes over time, unlike the aggregate method, which includes all past information. We see the MLH formulation is smoother than the sampling method, indicating that the most probable paths through the graph may be more important to consider than shortest paths. Finally, we note that our experiments used a relatively simple estimate of relationship strength for the edge probabilities in the network. In future work we will investigate alternative formulations of edge uncertainty.

Acknowledgements

This material is based in part upon work supported by the Intelligence Advanced Research Projects Activity (IARPA) via Air Force Research Laboratory contract number FA8650-10-C-7060. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon. 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 University Frederick N. Andrews Fellowship.

References

Bonacich, P. 1987. Power and Centrality: A Family of Measures. *The American Journal of Sociology* 92(5):1170–1182.

Boragatti, S. P. 2005. Centrality and Network flow. *Social Networks* (27):55–71.

Brandes, U. 2001. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology* 25:163–177.

Elandt-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, Vol. 17, No. 4 (Jul. - Aug., 1969), pp. 583-599.*

Freeman, L. C. 1977. A set of measures of centrality based on betweenness. *Sociometry* 40(1):35–41.

Goldenberg, J.; Libai, B.; and Muller, E. 2001. Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth. *Marketing Letters*.

Hua, M., and Pei, J. 2010. Probabilistic path queries in road networks: traffic uncertainty aware path selection. In *EDBT*, 347–358.

Kossinets, G.; Kleinberg, J.; and Watts, D. 2008. The structure of information pathways in a social communication network. In *KDD '08*, 435–443.

Leskovec, J.; Kleinberg, J.; and Faloutsos, C. 2005. Graphs over time: Densification laws, shrinking diameters and possible explanations. In *In KDD*, 177–187.

Marks, R. Enron timeline. *http://www.agsm.edu.au/bobm/teaching/BE/Enron/timeline.html*.

Newman, M. J. 2005. A measure of betweenness centrality based on random walks. *Social Networks* 27(1):39 – 54.

Onnela, J.-P.; Saramki, J.; Hyvnen, J.; Szab, G.; Lazer, D.; Kaski, K.; Kertsz, J.; and Barabsi, A.-L. 2007. Structure and tie strengths in mobile communication networks. *Proc Natl Acad Sci U S A* 104(18):7332–6.

Potamias, M.; Bonchi, F.; Gionis, A.; and Kollios, G. 2009. Nearest-neighbor queries in probabilistic graphs.

Raghavan, A.; Kranhold, K.; and Barrionuevo, A. Full speed ahead: How enron bosses created a culture of pushing limits. *http://academic.udayton.edu/lawrenceulrich/EnronBossesCreatingCulture.htm*.

Shetty, J., and Adibi, J. 2004. The enron email dataset database schema and brief statistical report.

Tang, J.; Musolesi, M.; Mascolo, C.; Latora, V.; and Nicosia, V. 2010. Analysing information flows and key mediators through temporal centrality metrics. In *SNS'10*.

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*.