Link Prediction Based on Graph Topology: The Predictive Value of the Generalized Clustering Coefficient

Zan Huang
Department of Supply Chain and Information Systems
Pennsylvania State University
419 Business Building
University Park, PA, 16802

zanhuang@psu.edu

ABSTRACT

Predicting linkages among data objects is a fundamental data mining task in various application domains, including recommender systems, information retrieval, automatic Web hyperlink generation, record linkage, and communication surveillance. In many contexts link prediction is entirely based on the linkage information itself (a prominent example is the collaborative filtering recommendation). Link-structure based link prediction is closely related to a parallel and almost separate stream of research on topological modeling of large-scale graphs. Graph topological modeling builds on random graph theory to find parsimonious graph generation models reproducing empirical topological measures that summarize the global structure of a graph, such as clustering coefficient, average path length, and degree distribution. These well-studied topological measures and graph generation models have direct implications on link prediction. This paper represents initial efforts to explore the connection between link prediction and graph topology. The focus is exclusively on the predictive value of the clustering coefficient measure. The standard clustering coefficient measure is generalized to capture higher-order clustering tendencies. The proposed framework consists of a cycle formation link probability model, a procedure for estimating model parameters based on the generalized clustering coefficients, and model-based link prediction generation. Using the Enron email dataset we demonstrate that the proposed cycle formation model corresponded closely with the actual link probabilities and the link prediction algorithm based on this model outperformed existing algorithms.

Categories and Subject Descriptors

G.3.3 [**Probability and Statistics**]: Statistical computing; G.2.3 [**Discrete Mathematics**]: Application

General Terms

Algorithms, Measurement.

Keywords

Link prediction, graph topological structure, clustering coefficient

1. INTRODUCTION

Many data mining tasks involve (sometimes implicitly) prediction of linkages among data objects. Examples of explicit link prediction problems include automatic Web hyperlink creation, genetic or protein-protein interactions prediction, and the record linkage problem. Other well-studied problems can be viewed as a link prediction problem once the data are rendered with a graph/network representation. Such examples are abundant. Information Retrieval can be viewed as dealing with prediction of links between words and documents within a word-document bipartite graph representing word occurrence. Recommender systems can be viewed as services predicting links between users and items within a user-item bipartite graph representing preferences or purchases.

While computational methods developed for these problems of different forms all essentially deal with the same fundamental problem of predicting links in network or graph data, link prediction has recently received substantial interest as a generic data mining task in the relational learning field [9]. Relational learning (or multi-relational learning) deals with richly structured data, which may be described by a relational database or using relational or first-order logic. Objects of multiple types can be linked with each other. The structure of the links or dependencies among data objects is the key data pattern specially exploited by relational learning methods. Various relational learning methods have been developed to predict the existence of potential links within a relational dataset that typically consists of observed linkages among data objects and attributes of the data objects.

Link structure has long been the focus of study for fields outside of data mining. Graph theory is essentially the study of link structure, but in an *abstract* graph/network where vertices are abstract data objects without descriptive attributes and edges are abstract links. Standard graph theory typically studies the properties of small-scale graphs. On the other hand, graph topological modeling based on the random graph theory, which have seen a surge in recent interest and a wide range of applications, investigates properties of large-scale real graphs such as the Internet, WWW, citation and co-authorship networks, and genetic regulatory networks. The primary focus here is on characterizing global patterns of the link structure, or graph topological feature, and developing parsimonious graph

generation models that can reproduce such features to describe the significant mechanisms governing the structure of the graph.

Although in general a complete relational data consisting of both link structure and descriptive attributes of data objects can be exploited, many link prediction methods are developed exclusively for exploiting the link structure itself. A prominent example is the collaborative filtering approach for recommender systems [22]. Under such an approach, attributes of the users and items are deliberatively ignored and only the links representing preferences or purchases are used for making recommendations (or potential links between users and items). When only the link structure is employed, the link prediction problem is formulated upon an abstract graph. Here we see a rarely-explored connection between the link prediction problem and the graph topological modeling. For example, a key question posed by graph topological modeling is whether or not a given large-scale graph is random. This question is answered by comparing the observed topological properties with the expectation from a random graph generation model. If we determine a given graph to be consistent with one generated from a random graph model, searching for data patterns for link prediction would be meaningless as no such pattern should exist. Beyond the basic randomness test, the well-studied topological measures can be viewed as a form of link data summary statistics and be employed directly to perform link prediction tasks in addition to building graph generation models explaining the graph data. After all, model building and prediction are rarely separable in statistical inference.

This paper focuses on analyzing the predictive value of one particular graph topological measure called clustering coefficient. Clustering coefficient describes the tendency to form clusters (fully connected subgraphs) in a graph. A typical clustering coefficient describes the probability for a connected triple to form triangles. We formalize the notion of generalized clustering coefficients to describe the formation of longer cycles. A link prediction framework based on the analysis of generalized clustering coefficients is then proposed, consisting of a cycle formation link probability model, a procedure for estimating model parameters based on the generalized clustering coefficients of a given graph, and model-based link prediction generation. An experimental study based on real-world graph data was conducted to demonstrate the proposed framework and its effectiveness.

The remainder of the paper proceeds as follows. Section 2 reviews relevant literature on link prediction, relational learning, and graph topological modeling. Section 3 introduces generalized clustering coefficient and describes the link prediction framework based on clustering coefficient analysis. Section 4 presents the experimental study. Section 5 concludes the paper and points out limitations and future directions.

2. LITERATURE REVIEW

2.1 The Link Prediction Problem

Prediction of links is the modeling focus of many well-studied problems. In many contexts, the link structure itself is the critical data pattern exploited for prediction. A wide range of problems can be viewed as prediction of links based on the observed link structure, including information retrieval [25] (predicting query-document links based on a document-word network), collaborative filtering recommendation [22] (predicting user-item

links based on a user-item interaction matrix), record linkage problem [30] (predicting links among records with same identity), and protein/genetic interaction modeling [12] (predicting underlying protein/genetic interactions based on interaction networks observed from experiments). Many algorithms developed for these problems also work for the generic link prediction problem. On the other hand, advances in the link prediction problem would have potential implications in these different application domains.

Recently, the link prediction problem has been formulated as a generic data mining task within the field of relational learning. *Relational learning* or *multirelational learning* [6] extends standard data mining that learns from attributes of independent entities stored in a single database table to extract patterns from multiple related tables. Link structure is exploited by relational learning methods to enhance the performance of various well-known data mining tasks such as classification and clustering [9]. Link prediction, where link structure itself is the target of prediction, has also become an important task of relational learning.

Various relational learning methods were proposed for link prediction, typically exploiting both the link structure itself and the rich descriptive attributes of data objects. *Probabilistic relational models* (PRMs) [16] are the main formal approach that has been developed for relational learning, which extends *Bayesian networks* to the relational domain. Getoor et al. [10] introduced the concept of structure uncertainty and extends the PRMs to model and predict link existence. Huang et al. extended the PRMs framework further to work specifically for link prediction in a recommendation context [15]. Several studies also employed other relational probabilistic graphical models (e.g., *relational Markov network*) for the link prediction problem [5, 23, 27]. Propescul and Ungar [21] applied the structural logistic regression model for building link prediction models.

Relational learning link prediction models typically assume a richly structured relational data environment, where type information regarding the data objects and linkages, regular descriptive attributes, link structure, and potentially descriptive attributes of the links are available. These models can be directly applied to abstract graphs where link structure is the only source of predictive data patterns. However, many of these models might not able to capture the well-established data patterns in large graphs such as paths, cycles, and flows.

The link prediction problem on abstract graphs (networks of no vertex and edge attributes) is the focus of our study. Liben-Nowell and Kleinberg [17] studied the abstract graph link prediction problem for social networks with an abstract graph representation. They were specifically looking at academic co-authorship networks. They investigated the relative effectiveness of network proximity measures adapted from graph theory, computer science, and social science and confirmed for social networks the power of prediction methods based purely on the graph structure. Many other algorithms developed in other fields that deal with problems with implicit abstract graph representation are also suitable for abstract graph link prediction, such as the collaborative filtering recommendation algorithms [14].

2.2 Graph Topological Modeling

Abstract graphs are the exclusive focus for the field of graph theory, for which properties associated with concepts such as paths, cycles, and flows within relatively small-scale graphs are the primary interest. Recent surge of application of random graph theory and topological graph modeling in a wide variety of scientific, engineering, and social domains [1, 19] has switched focus to analyzing structural features of large-scale complex graphs, which is deeply associated with the idea of discovering patterns from large-scale datasets in data mining.

Random graph modeling research exploits a graph representation of complex systems with a focus on its topological characteristics. Its main research objective is to capture the mechanisms that determine the network topology of a particular system. The key assumption is that the fundamental mechanism that governs the generation of relationships among components of a system leaves certain identifiable traits in the resulting network topology. Thus, a simple graph generation model that can reproduce similar topological features of the real network may bring important insights to the understanding of the actual mechanism that governs the real system.

Many recent studies show that real-world networks demonstrate surprisingly consistent topological characteristics across different domains [1]. Three major concepts related to these topological features are "small world," "clustering," and "scale-free" phenomena, which involve three basic topological measures: the *average path length*, *clustering coefficient*, and *degree distribution*. The average path length measure is defined as the average distance between any pair of nodes. The *degree* of a vertex in a graph is the number of edges incident on that vertex. In this paper we focus on analyzing the clustering coefficient measure, which we introduce in detail below.

Many real-world networks show an inherent tendency to cluster. Such a tendency is quantified by the clustering coefficient measure [20, 29]. We adopt the Newman definition:

$$C = \frac{3 \times (\text{number of triangles in the graph})}{\text{number of connected triples}} \quad (1)$$

where a triangle is a set of three vertices each of which is connected to both of the others, and a connected triple is three vertices *x-y-z*, with both vertices *x* and *z* connected with *y* (note that *x-y-z* and *z-y-x* are considered the same connected triple). The clustering coefficient *C* is strictly bounded between 0 and 1 and measures the extent to which being a neighbor is a transitive property.

Various graph generation models have been proposed, ranging from the classic purely random ER model [7], in which the generation of the graphs is conditional only on the size of the graph and the vertex connection probability, to various hybrid random graph generation models that incorporate certain non-random principle to reproduce empirically observed topological characteristics (e.g., [29, 2]).

Graph topological measures can be viewed as a form of summary statistics for graph data. Just like the correlation coefficient summarizes a sample of two random variables, topological measures summarize graph data patterns relevant to building a graph generation model explaining the link occurrences in the observed graph. These measures should provide valuable information for understanding of the link structure and prediction of future or unobserved links.

3. LINK PREDICTION BASED ON GENERALIZED CLUSTERING COEFFICIENTS

3.1 Generalized Clustering Coefficient

In this paper, we focus on analyzing the predictive power of clustering coefficient. The tendency to form clusters is an important aspect of graph structural patterns. The standard clustering coefficient measures the probability for a connected triple to form a triangle. Such a data pattern naturally exhibits a predictive power. Given a graph with a large clustering coefficient, links that would lead to many new triangles would naturally be good candidates of future or missing links.

Although most of the existing literature on graph topological modeling has focused on the average path length, clustering coefficient, and degree distribution measures, no formal study investigates whether these measures are sufficient in characterizing graph generation process of real-world graphs. It has been shown that the standard clustering coefficient does not fully capture the clustering mechanism in real graphs [11]. Several studies have analyzed longer cycles in graphs and looked at higher order clustering coefficients that measure tendency in a graph to form longer cycles (e.g., [8, 11]).

We now introduce the formal notations involving a graph and define the generalized clustering coefficient. To be consistent with the graph studied by the existing topological modeling literature, we limit our focus on abstract graphs with undirected links. Let $G = (V, E)$ be a finite undirected graph without multiple edges or self-loops. $V = (1, 2, \ldots, N)$ is the list of vertices of G, and $E = (e_1, e_2, \ldots, e_M)$ is the list of edges of G, each e_s corresponds to a sequence of two vertices (i, j). In this paper, we use edge and link interchangeably. Since we are looking at undirected graphs, if (i, j) is an edge of G, (j, i) will also be an edge of G. A *path* of length k in G is a nonempty sequence of vertices $p = (v_0, v_1, \ldots, v_k)$ such that (v_i, v_{i+1}) is an edge of G for all i, $0 \leq i \leq k - 1$. A *cycle* of length k in G is a nonempty sequence of vertices $p = (v_0, v_1, \ldots, v_k, v_0)$ such that (v_i, v_{i+1}) is an edge of G for all i, $0 \leq i \leq k - 1$, and (v_k, v_0) is an edge of G. P_{ijk} denotes the set of paths of length k starting at i and ending at j. We denote the number of such paths as $|P_{ijk}|$.

A generalized clustering coefficient $C(k)$ of degree k is defined as:

$$C(k) = \frac{\text{number of cycles of length } k \text{ in the graph}}{\text{number of paths of length } k} \quad (2)$$

When $k = 2$, (2) reduces to (1) as the number of paths of length 2 doubles the number of connected triples and the number of cycles of length 2 is six times the number of triangles in the graph. When computing $C(k)$ for a undirected graph, we can avoid redundant counting by only counting paths with smaller-index vertex as the starting vertex and checking for each path whether the starting vertex and ending vertex form an edge in the graph (if so a cycle is counted). The obtained numbers for the cycles and paths will be exactly half of the total number of cycles and paths, and their ratio gives the clustering coefficient measure according to (2). To compute the generalized clustering coefficients, Rubin's algorithm for enumerating all simple paths [24] can be employed.

Rubin's algorithm uses $O(N^3)$ matrix operations to find all paths of different lengths in a graph. The algorithm can be customized to find all paths of length up to k for our purpose. Figure 1 presents an example graph illustrating clustering coefficients of different degrees.
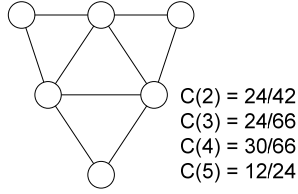

C(2) = 24/42
C(3) = 24/66
C(4) = 30/66
C(5) = 12/24

Figure 1. An example illustrating generalized clustering coefficients.

3.2 A Link Probability Model Based on Cycle Formation

Generalized clustering coefficients capture important graph topological characteristics. From the link prediction perspective, these summary statistics regarding the graph structure directly correspond to the conditional probability of observing cycles of certain length given the same-length paths. In other words, the clustering coefficients describe the correlation between cycles and paths.

Previous link prediction studies have tried to exploit the cycle formation tendency observed in real graphs in a qualitative manner. These previous efforts can be broadly categorized into *local* and *global* methods. Under the local approach (e.g., the common neighbors, Jaccard's coefficient, and Adamic/Adar measure in [17] and standard user-based and item-based collaborative filtering algorithms [4]), link occurrence probability is heuristically set to positively correlate with the number of common neighbors (or the number of cycles of length 2 that would be formed if the link at question existed). The global approach (e.g., the spreading activation approach [13] and the Katz, hitting time, PageRank, and variants in [17]) explores longer paths between a given pair of vertices and essentially relate the link occurrence probability to the sum of total number of paths between the two vertices weighted by the path length. These previous methods are *heuristic* in nature and rely implicitly on the assumed high tendency for paths of different lengths to form cycles.

The local approach methods are qualitatively supported by many recent empirical findings that real networks across different domains share the common property of high clustering tendency when measured by the standard clustering coefficient measure ($C(2)$). Previous approach essentially attempt to predict links that will maximize the clustering coefficient of the resulting network, which is implicitly assuming the network is on the evolution path to higher clustering tendency and ultimately to a fully connected network. In reality, many networks across different domains have exhibited stationary level of clustering tendency independent of the size of the network (see Figure 9 in [1]). Under this stationary assumption, the exact value of observed clustering coefficient should be incorporated to design the link prediction algorithms, such that the predicted links would grow the network with consistent clustering coefficient over time. To the best of our knowledge, no formal link probability model has been

investigated previously based on the underlying clustering coefficient (propensity for length-2 paths to become triangles).

The global methods are natural extensions of the local methods. If links are more likely to occur between nodes connected by short paths (of length 2), they might be more likely to occur between nodes connected by longer paths as well. However, these methods are not justified by even qualitative empirical observations as the empirical findings on clustering coefficients of higher degrees ($k > 2$) are quite limited. The lack of understanding on real networks' higher-degree clustering coefficient properties poses challenges on the heuristic link prediction algorithms that assume positive correlation between paths and cycles of large length.

In this paper, we propose a link probability model based on cycle formation that formally relates to generalized clustering coefficients. In this model, the occurrence probability of a particular link is determined by the number of cycles (of different lengths) that will be formed by adding this link. A fundamental assumption of our framework is the stationary property of the degree of clustering of the network. The model parameters, corresponding to the prevalence of mechanisms such as length-k cycle formation, can be estimated given a series of generalized clustering coefficients (of different degrees). Assuming the clustering stationarity, this model is then used to predict future or unobserved links in the network.

Formally, a *cycle formation model* of degree k ($k \geq 1$), denoted as $CF(k)$, treats the link occurrence probability as governed by t link generation mechanisms, $g(1), \ldots, g(k)$, each described by a single parameter, $c_1, \ldots, c_k$. Among these mechanisms, $g(1)$ is a dummy mechanism that corresponds to a random link generation process and c_1 corresponds to the probability for a link to occur within a purely random model. Therefore $CF(1)$ reduces back to a random link probability model. Other mechanisms $g(k)$'s ($k > 1$) correspond to link probability determined by the length-k paths associated with the vertex pair. The parameter $c_k = \Pr((i, j) \in E \mid |P_{ijk}| = 1)$ describes the conditional probability of a length-k to become a length-k cycle.
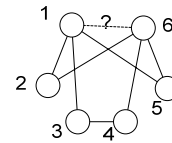

Figure 2. An example illustrating the cycle formation link probability model.

For example, assume a cycle formation model of degree 3 ($CF(3)$), the occurrence probability of the potential link (1, 6) in Figure 2 is governed by three mechanisms: the random link occurrence $g(1)$, length-2 cycle formation mechanism $g(2)$, and length-4 cycle formation mechanism $g(3)$.

In order to make our link probability model complete, we need to decide how multiple link generation mechanisms should be combined to derive the total link occurrence probability. We will start with integration of multiple paths of the same length. Take the two length-2 paths between 1 and 6 in Figure 2 as an example, it is key in our context to realize that path 1−2−6 and 1−5−6 cannot be treated as two independent determinants of the occurrence probability of (1, 6). Depending on the occurrence of edge (1, 6), both paths either remain paths or become cycles

together. Therefore, given $\Pr((i, j) \in E \mid |P_{ij2}| = 1) = c_1$, $\Pr((i, j) \in E \mid |P_{ij2}| = 2) = c_1^2/(c_1^2+(1-c_1)^2)$. To generalize:

$$\Pr((i, j) \in E \mid |P_{ijk}| = m) = c_k^m/(c_k^m+(1-c_k)^m), \ k > 1 \qquad (3)$$

Similarly, we combine the effects of multiple mechanisms to form the total link occurrence probability under the *cycle formation model of degree k* (*CF(k)*), denoted as $P_{m_2,...,m_k}$, as follows:

$$P_{m_2,...,m_k} = \Pr((i, j) \in E \mid\mid P_{ij2} \mid= m_2,...,\mid P_{ijk} \mid= m_k) =$$
$$c_1 c_2^{m_2}...c_k^{m_k} /(c_1 c_2^{m_2}...c_k^{m_k} + (1-c_1)(1-c_2)^{m_2}...(1-c_k)^{m_k}) \qquad (4)$$

Our model accounts for a variety of detailed link occurrence patterns based on cycle formation. When $c_k = 0.5$, length-k paths do not have any effect on the link probability for the vertex pair, while a c_k greater (smaller) than 0.5 indicates that length-k paths have positive (negative) effects on the linking probability.

As our link probability model was motivated by the generalized clustering coefficients, it is important to relate the model back to the clustering coefficient measures and eventually based on these empirical measures to estimate the model parameters. In this paper we limit the discussion on cycle formation models up to degree of 4. Table 1 shows the details for deriving expected clustering coefficients of degrees 2, 3, and 4 based on the link probability model in (4). Higher-degree models can be derived in a similar manner.

Table 1. Expected clustering coefficients of degrees 2, 3 and 4 based on the cycle formation model $(E[C(k)] = f(c_1,c_2,...,c_k) = \sum_i \#(G_i)\Pr(G_i)\Pr((1,k+1) \in E \mid G_i))$

k	i	Graph Pattern (G_i)	$\#(G_i)$	$\Pr(G_i)$	$\Pr((1,k+1) \in E \mid G_i)$
2	1		1	1	$P_{1,0,0}$
3	1		1	$(1-P_3)^2$	$P_{0,1,0}$
	2		2	$P_3(1-P_3)$	$P_{1,1,0}$
	3		1	P_3^2	$P_{2,1,0}$
4	1		1	$(1-P_4)^2(1-P_3)^3$	$P_{0,0,1}$
	2		3	$(1-P_4)^2(1-P_3)^2 P_3$	$P_{0,1,1}$
	3		2	$P_4(1-P_4)(1-P_3)^3$	$P_{1,0,1}$
	4		1	$(1-P_4)^2(1-P_3)P_3^2$	$P_{1,2,1}$
	5		2	$(1-P_4)^2(1-P_3)P_3^2$	$P_{0,2,1}$
	6		1	$P_4^2(1-P_3)^3$	$P_{2,0,1}$
	7		6	$(1-P_4)P_4(1-P_3)^2 P_3$	$P_{1,1,1}$
	8		1	$(1-P_4)^2 P_3^3$	$P_{1,3,1}$
	9		2	$(1-P_4)P_4(1-P_3)P_3^2$	$P_{2,2,1}$
	10		4	$(1-P_4)P_4(1-P_3)P_3^2$	$P_{1,2,1}$
	11		3	$P_4^2 P_3(1-P_3)^2$	$P_{2,1,1}$
	12		2	$P_4^2 P_3^2(1-P_3)$	$P_{2,2,1}$
	13		1	$P_4^2 P_3^2(1-P_3)$	$P_{3,2,1}$
	14		2	$(1-P_4)P_4 P_3^3$	$P_{2,3,1}$
	15		1	$P_4^2 P_3^3$	$P_{3,3,1}$

Table 1 shows all possible graph patterns relevant to a given path of length k ($p = (1,2,..., k+1)$ in this context). For each possible graph pattern (G_i), the number of subgraphs corresponding to this pattern ($\#(G_i)$), the probability for one of such subgraphs to occur ($\Pr(G_i)$), and the probability for the edge $(1, k+1)$ to occur conditional on G_i ($\Pr((1, k+1) \in E \mid G_i)$) are shown. $\Pr((1, k+1) \in E \mid G_i)$ is derived from formula (4) with $|P_{ijk}|$'s computed from the specific graph pattern G_i. The total probability for observing the link $(1, k+1)$ conditional on a path $p = (1, 2, ..., k+1)$ is given by

$$f(c_1,c_2,...,c_k) = \sum_i \#(G_i)\Pr(G_i) \ \Pr((1,k+1) \in E \mid G_i). \qquad (5)$$

This probability is the theoretical prediction of the expected clustering coefficient of degree k ($E[C(k)]$) based on our cycle formation model.

3.3 Estimating the Model Parameters and Performing Link Prediction

We propose an iterative procedure for estimating the parameters of the cycle formation model based on the generalized clustering coefficients. The estimation procedure is described below in Figure 3.

At Step 2, a random graph with the degree distribution of G can be viewed as generated from the $CF(1)$ model. Theoretically, $C(k)$ = c_1, $\forall$ $k > 2$ under $CF(1)$ and c_1 can be computed as the average linking probability of the ER model [7] $2M/N(N-1)$ for very large graphs. We found that random graphs of hundreds of vertex and edges with degree distributions consistent with real networks typically exhibit $C(k)$'s that deviate from the linking probability of the ER model. For different k, $C(k)$ also slightly varied. For our purpose, we adopted a numeric approach to generate random graphs with consistent degree distribution as the input graph G (using the switching algorithm in [18]) and took the average, $C(2)_{rand}$, as the estimate of c_1 for the cycle formation model.

Parameter Estimation for Cycle Formation Model $CF(k)$:
Input: $G = (V, E)$
Output: $c_1, c_2, \ldots, c_k$
1. Compute generalized clustering coefficients $C(2), \ldots, C(k)$
2. Compute the connecting probability under a random graph with the degree distribution of G as c_1
3. Set $c_2 = (1-c_1)C(2)/(c_1-2c_1C(2)+C(2))^*$
4. Set $c_i = 0.5$, $i = 3, \ldots, k$
5. Repeat for i = 3, …, k
5.1. $c_i = \underset{c_i'}{\arg\min}(|\,C(i) - f(c_1, \ldots, c_i', \ldots, c_k)\,|)$ [†]

Figure 3. Procedure for estimating parameters of the cycle formation model.

Our parameter estimation procedure builds directly upon the cycle formation model introduced in Section 3.2. Intuitively, we start with a pure random model conditional on the degree distribution (with no meaningful cycle formation mechanism at all) and obtain the cycle formation probability c_1. We then compare the observed $C(2)$ and c_1. The part of $C(2)$ that cannot be explained by c_1 has to be due to a meaningful length-2 cycle formation mechanism. Thus the estimator of c_2 can be derived. We then proceed with comparison between observed $C(3)$ with the expected $C(3)$ under $CF(2)$ to obtain the estimate of c_3. This process continues until reaching the degree of the model. The key to our approach is that $C(k)$ is a function of $c_1, \ldots, c_k$, and is independent of $c_{k'}$, $k' > k$. With the estimated parameters of the cycle formation model, we then use formula (4) to derive the link probabilities.

4. EXPERIMENTAL STUDY
4.1 Data
We used the Enron email dataset to evaluate our proposed cycle formation link probability model and the corresponding link prediction algorithm. The Enron email corpus is a large-scale email collection from a real organization over the course covering a 3.5 years period. We used a pre-processed version of the dataset provided by Jitesh Shetty and Jafar Adibi [26] (data available at ftp://ftp.isi.edu/sims/philpot/data/enron-mysqldump.sql.gz). This dataset contains 252,759 emails from 151 Enron employees, mainly senior managers. In our study we have focused on emails sent from *and* to these 151 people.

The final email collection we analyzed contained 40,489 emails during May 11th of 1999 to June 21st of 2002. We have decided

---

* Derived from $C(2) = c_1c_2/(c_1c_2 + (1-c_1)(1-c_2))$.
[†] Typical numerical methods can be used to find c_i that gives $f(c_1, c_2, \ldots, c_k)$ within $[C(i) - \varepsilon, C(i) + \varepsilon]$ for specified small ε.

---

to perform the link prediction analysis on the monthly email graphs in 2001. In this study, an email graph is an undirected and unweighted graph with edges connecting senders and recipients of emails during the corresponding time periods. The semantics of an edge (a, b) in such a graph is that there has been at least one email communication between a and b (either a sending at least one email with recipients including b or b sending at least one email with recipients including a). For month t in 2001, we used the emails in the previous three months ($t-3$, $t-2$, $t-1$) to form the background graph G_{tb}. This background graph is the input for prediction of email links in G_t. Like most of the link prediction problems, we are interested in predicting the occurrence of links in G_t that did not already appeared in G_{tb}. Table 2 shows the number of links in G_{tb} and generalized clustering coefficient measures up to degree 4. We can observe from the table that across different months in 2001, the background email graphs maintained relatively stable clustering coefficients despite the constantly increasing number of links.

In addition to the generalized clustering coefficients, the connecting probability under random graphs with the same degree distribution as the background email graphs are also needed as the input for the cycle formation model parameter estimation procedure. As described in Section 3.3, we used the numerically derived $C(2)_{rand}$ as the estimate of c_1. In our study, $C(2)_{rand}$ for G_t were averaged from 20 random graphs with the same degree distribution as G_t.

Table 2. Background graph measures: number of links and generalized clustering coefficients

Month (t)	Links in G_{tb}	$C(2)$	$C(3)$	$C(4)$	$C(2)_{rand}$
1	293	0.33096	0.23823	0.14936	0.08640
2	320	0.32601	0.23311	0.14551	0.09323
3	322	0.32209	0.22479	0.14167	0.10626
4	319	0.30438	0.23183	0.14594	0.11809
5	372	0.33735	0.24507	0.16657	0.10776
6	466	0.25617	0.18714	0.13372	0.11647
7	500	0.24860	0.17575	0.12921	0.12693
8	521	0.26801	0.19199	0.14545	0.11411
9	564	0.30663	0.23026	0.18632	0.13281
10	580	0.32155	0.24162	0.19575	0.12692
11	763	0.33394	0.25565	0.20934	0.14318
12	763	0.36168	0.26508	0.20998	0.13900

4.2 Cycle Formation Model and Parameter Estimation
A basic test on the validity of the proposed cycle formation model is to verify the accuracy of the theoretical expected clustering coefficients. In our current study, we generated random graphs that have the same degree distribution and standard clustering coefficient ($C(2)$ in our context) as the 12 real background email graphs. We adopted Volz's algorithm for random graphs with tunable degree distribution and clustering [28] to generate these graphs. In principle, these random graphs can be viewed as generated from the cycle formation model $CF(2)$. For each graph G_t, c_1 and c_2 were estimated following the procedure described in Section 3.3. The expected $C(3)$ and $C(4)$ were computed assuming a $CF(2)$ model and then compared with the actual measures of the randomly generated graphs. Figure 4 shows the comparison between the expected $C(3)$ and $C(4)$ under $CF(2)$ based on the cycle formation model with the actual values of $C(3)$

and $C(4)$ of the randomly generated graphs that are consistent with $CF(2)$. For the randomly generated graphs, we show the mean and 95% confidence interval of the measures based on samples of 10 random graphs. Even with only 10 random graphs as the samples, we observe from the figure that the theoretical predictions of the higher-degree clustering coefficients were generally consistent with the randomly generated graphs.
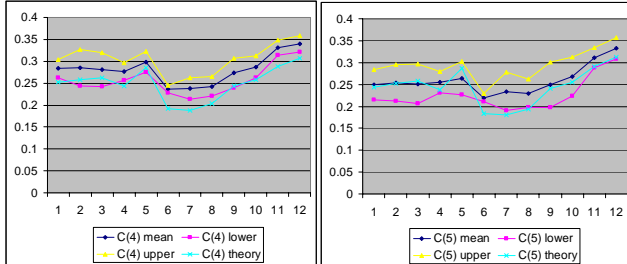

Figure 4. Theoretical and simulated values of $C(3)$ and $C(4)$ under $CF(2)$.

Following the parameter estimation procedure described in Section 3.3, we estimated the parameters for the complete $CF(4)$ models for the 12 background email graphs. The parameter estimates are shown in Table 3. Notice that c_1 is identical to $C(2)_{rand}$ in Table 2 as previously discussed in Section 3.3. From these estimates, we observe that over the 12 months period, the Enron email graphs have exhibited consistent cycle formation patterns. The prominent patterns (parameter estimates largely differ from 0.5) were the positive effect of length-2 cycle formation and the negative effect of length-3 cycle formation. Length-4 cycle formation had relatively minor effect on link probability as for most months c_4 was close to 0.5. The negative effect of length-3 cycles was especially interesting. Previous link prediction approaches typically assume unidirectional effect of formation of cycles of different lengths and would misinterpret the information embedded in the length-3 paths.

Table 3. Parameter estimates for the cycle formation model

Month (t)	c_1	c_2	c_3	c_4
1	0.08640	0.83951	0.26563	0.55039
2	0.09323	0.82469	0.26211	0.55068
3	0.10626	0.79985	0.21367	0.58877
4	0.11809	0.76568	0.40625	0.43164
5	0.10776	0.80826	0.24648	0.59258
6	0.11647	0.72320	0.40781	0.47344
7	0.12693	0.69473	0.22656	0.57383
8	0.11411	0.73975	0.34844	0.53281
9	0.13281	0.74277	0.32695	0.57334
10	0.12692	0.76527	0.31445	0.58857
11	0.14318	0.75003	0.29883	0.60439
12	0.13900	0.77825	0.18906	0.69990

4.3 Link Prediction Performance

In this study, we are interested in predicting email links that did not appear in the background email graph. Table 4 shows the number of new email links for each of the 12 months. To evaluate the link prediction performance, we construct a Receiver Operating Characteristics (ROC)-style curve with x-axis and y-axis as the percent of total possible new links selected and the percent of actual new links that are in the selected links. The area

under curve (AUC) measure [3] is reported for assessing the link prediction performance.

For comparison purpose, we also reported the link prediction performance of three representative existing link prediction algorithms: (a) the *preferential attachment* (*PA*) algorithm [17] that was motivated by the preferential attachment model [2] and relate the link probability with the product of the degrees of the two vertices; (b) the *spreading activation* (*SA*) algorithm [13] that explores the ensemble of paths connecting the vertex pair of all lengths and heuristically relate larger number of paths of different lengths with higher link probability (this algorithm is in essence similar to Katz, hitting time, PageRank and variants in [17]); (c) the *generative model* (*GM*) algorithm [14] that introduces latent email types and employs the *Expectation Maximization* algorithm to estimate the probability of each people to be associated with these latent types as senders and recipients. Both the spreading activation and generative model algorithms have demonstrated competitive performance over other algorithms in many link prediction tasks such as the collaborative filtering task.

Based on the analysis in the previous section, we decided to use the $CF(3)$ model as the parameter c_4 does not seem to deviate from 0.5 significantly. Using the parameter estimates for the $CF(3)$ model we computed the link probability scores as described in Section 3.3. The AUC measures of our proposed algorithm as well as the three benchmark algorithms are shown in Table 4. The $CF(3)$ algorithm achieved the highest AUC measure for the first 9 months and achieved the second-based performance for the remaining 3 months. There was clearly no preferential attachment phenomenon in our data. The PA algorithm was worse than a random predictor for most months. The SA and GM algorithms had comparable performances that were substantially worse than our proposed algorithm.

Table 4. Number of new links and AUC measures for the cycle formation and benchmark link prediction algorithms

Month (t)	New Links in G_t	$CF(3)$	PA	SA	GM
1	46	**0.76764**	0.52091	0.70977	0.67237
2	46	**0.76748**	0.48449	0.67891	0.69189
3	46	**0.81385**	0.44235	0.72598	0.73472
4	62	**0.82106**	0.45057	0.72143	0.75518
5	95	**0.73883**	0.45141	0.63052	0.69805
6	72	**0.71737**	0.41585	0.67084	0.70634
7	88	**0.73977**	0.48104	0.70304	0.69231
8	168	**0.74056**	0.60122	0.69806	0.65201
9	90	**0.72065**	0.46032	0.69501	0.70263
10	213	0.74384	0.48877	0.69840	**0.74500**
11	123	0.74860	0.47632	0.72477	**0.75116**
12	56	0.70018	0.44078	0.69573	**0.76273**

5. CONCLUSIONS AND FUTURE DIRECTIONS

In this paper we explore the connection between two largely separate fields of link prediction and graph topology modeling. The key idea is that the well-studied topological measures in effect serve as summary statistics describing the link occurrences in graphs and carry valuable information for building model-based link prediction algorithms. In this study, we focus on analyzing generalized clustering coefficients and their prediction value for link prediction. We proposed a cycle formation model

Sthat relates the occurrence probability a link with its ability for form cycles of different lengths. The parameters of this model can be estimated given the generalized clustering coefficients of the graph. Using the Enron email dataset we have verified that the cycle formation model was able to capture closely the actual link probabilities and that the link prediction algorithm based on this model outperformed existing link prediction algorithms.

Our framework can be enhanced in several aspects. Although expected clustering coefficients for higher degrees can be derived in a similar manner as shown in Table 1, a compact formula that approximates the exact value would ease the application of our framework. Closed-form prediction of the expected clustering coefficients of random graphs of given degree distribution would give the estimate for c_1 without having to numerically generate samples of such random graphs. Conceptually, our proposed link prediction framework shares the flavor as an autoregressive regression model for time series analysis, with generalized clustering coefficients corresponding to autocorrelation between present value and past values and link generation mechanisms $g(i)$'s corresponding to past values as predictors of the present value. Accordingly, to make our proposed framework complete, stochastic component should be included in our model and confidence interval of the parameter estimates should be derived to perform complete statistical inference and model selection.

6. REFERENCES

[1]   Albert, R. and Barabasi, A.-L. Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74*. 47-97, 2002.

[2]   Barabasi, A.-L. and Albert, R. Emergence of scaling in random networks. *Science*, *286*. 509-512, 1999.

[3]   Bradley, A.P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30* (7). 1145-1159, 1997.

[4]   Deshpande, M. and Karypis, G. Item-based top-N recommendation algorithms. *ACM Transactions on Information Systems*, *22* (1). 143-177, 2004.

[5]   Domingos, P. and Richardson, M., Mining the network value of customers. in *Sventh ACM SIGKDD international conference on Knowledge discovery and data mining*, (San Francisco, California, 2001), 57 - 66.

[6]   Dzeroski, S. and Lavrac, N. *Relational Data Mining*. Springer-Varlag, Berlin, 2001.

[7]   Erdos, P. and Renyi, A. On random graphs. *Publicationes Mathematicae*, *6*. 290-297, 1959.

[8]   Fronczak, A., Holyst, J.A., Jedynak, M. and Sienkiewicz, J. Higher order clustering coefficients in Barabási-Albert networks. *Physica A*, *316* (1-4). 688-694, 2002.

[9]   Getoor, L. Link mining: A new data mining challenge. *SIGKDD Explorations*, *5* (1). 84-89, 2003.

[10] Getoor, L., Friedman, N., Koller, D. and Taskar, B. Learning probabilistic models of link structure. *Journal of Machine Learning Research*, *3*. 679-707, 2002.

[11] Gleiss, P.M., Stadler, P.F., Wagner, A. and Fell, D.A. Relevant cycles in chemical reaction network. *Advances in Complex Systems*, *4*. 207-226, 2001.

[12] Goldberg, D.S. and Roth, F.P., Assessing experimentally derived interactions in a small world. in *National Academy of Sciences*, (USA, 2003), 4372-4376.

[13] Huang, Z., Chen, H. and Zeng, D. Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. *ACM Transactions on Information Systems (TOIS)*, *22* (1). 116-142, 2004.

[14] Huang, Z., Zeng, D. and Chen, H. A comparative study of recommendation algorithms for e-commerce applications. *IEEE Intelligent Systems*, *forthcoming*, 2006.

[15] Huang, Z., Zeng, D. and Chen, H., A unified recommendation framework based on Probabilistic Relational Models. in *Fourteenth Annual Workshop on Information Technologies and Systems (WITS)*, (Washington, DC, 2004), 8-13.

[16] Koller, D. and Pfeffer, A., Probabilistic frame-based systems. in *Fifteenth Conference of the American Association for Artificial Intelligence*, (Madison, Wisconsin, 1998), 580-587.

[17] Liben-Nowell, D. and Kleinberg, J., The link prediction problem for social networks. in *12th International Conference on Information and Knowledge Management (CIKM)*, (New Orleans, LA, 2003), 556 - 559.

[18] Milo, R., Kashtan, N., Itzkovitz, S., Newman, M.E.J. and Alon, U. Uniform generation of random graphs with arbitrary degree sequences. *cond-mat/0312028*, 2003.

[19] Newman, M.E.J. The structure and function of complex networks. *SIAM Review*, *45* (2). 167-256, 2003.

[20] Newman, M.E.J., Strogatz, S.H. and Watts, D.J. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev.*, *E 64*. 026118, 2001.

[21] Popescul, A. and Ungar, L., Statistical relational learning for link prediction. in *Workshop on Learning Statistical Models from Relational Data at the International Joint Conference on Artificial Intelligence*, (Acapulco, Mexico, 2003).

[22] Resnick, P., Iacovou, N., Suchak, M., Bergstorm, P. and Riedl, J., GroupLens: An open architecture for collaborative filtering of netnews. in *ACM Conference on Computer-Supported Cooperative Work*, (1994), 175-186.

[23] Richardson, M. and Domingos, P., Mining knowledge-sharing sites for viral marketing. in *Eighth International Conference on Knowledge discovery and Data Mining (SIGKDD'02)*, (Edmonton, Alberta, Canada, 2002), ACM Press, 61-70.

[24] Rubin, F. Enumerating all simple paths in a graph. *IEEE Transactions on Circuits and Systems*, *25* (8). 641-642, 1978.

[25] Salton, G. *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer*. Addison Wesley, Reading, MA, 1989.

[26] Shetty, J. and Adibi, J. The Enron dataset database schema and brief statistical reprort, Information Sciences Institute, University of Sothern California, 2005.

[27] Taskar, B., Wong, M.-F., Abbeel, P. and Koller, D., Link prediction in relational data. in *Neural Information Processing Systems*, (2004).

[28] Volz, E. Random networks with tunable degree distribution and clustering. *Physical Review E*, *70* (056115), 2004.

[29] Watts, D.J. and Strogatz, S.H. Collective dynamics of small-world networks. *Nature*, *393*. 440-442, 1998.

[30] Winkler, W.E. Advanced methods for record linkage. *Technical Report, Statistical Research Division, U.S. Census Bureau*, 1994.