The Evolution of Scientific Paper Title Networks

Vahed Qazvinian
Department of EECS
University of Michigan
Ann Arbor, MI
vahed@umich.edu

Dragomir R. Radev
School of Information
Department of EECS
University of Michigan
Ann Arbor, MI
radev@umich.edu

Abstract

In spite of enormous previous efforts to model the growth of various networks, there have only been a few works that successfully describe the evolution of latent networks. In a latent network edges do not represent interactions between nodes, but show some proximity values. In this paper we analyze the structure and evolution of a specific type of latent networks over time by looking at a wide range of document similarity networks, in which scientific titles are nodes and their similarities are weighted edges. We use scientific papers as the corpora in order to determine the behavior of authors in choosing words for article titles. The aim of our work is to see whether term selection for titles depends on earlier published titles.

Introduction

Modeling the behavior of different networks has received great attention in the past decade. These models are based on the facts observed by looking at the network properties in an interval of equally spaced points in time. The evolution of a wide range of networks have been already modeled. These models describe the growth of citation networks (Leskovec, Kleinberg, & Faloutsos 2005), friendship networks (Jin, Girvan, & Newman 2001), online social networks (Kumar, Novak, & Tomkins 2006; Leskovec *et al.* 2008), and many others. In all of the mentioned networks, nodes are entities that interact with each other by making links, and the links between them show relationships. Therefore, not every pair of nodes is connected. These networks are usually unweighted, such as friendship networks, and sometimes directed as in the citation networks.

Another type of networks are those in which edges represent proximity or similarity values rather than relationships. These networks are fully connected, weighted, and symmetric (if the proximity measure is symmetric). Applying a cutoff equal to c on this network, and pruning the edges with values smaller than c will make it a regular binary network We refer to A, which results in an ensemble of different binary networks, as a *latent network*. Figure 1 shows a sample latent network created by the 11 sentences from the LexRank sample dataset used in (Erkan & Radev 2004) at

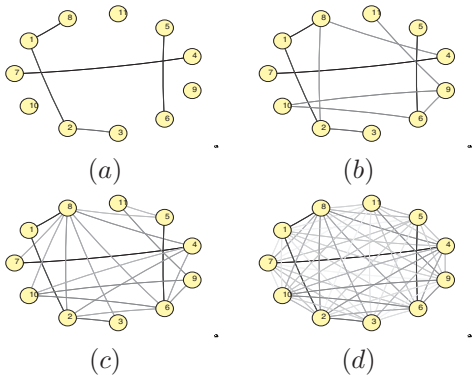


Figure 1: Lexical network for the 11-sentence sample at cutoff values (a) 0.20, (b) 0.15, (c) 0.10, and (d) 0.00.

different cutoff values. At each cutoff, c, the network consists of the edges with weights above c.

Most of the properties in regular networks are generalized to weighted ones, and therefore are applicable to latent networks.

Degree Newman (Newman 2004) defines the degree of a node in such a network as the sum of the weights of all links attached to it, $k_i = \sum_j \mathcal{L}_{ij}$.

Shortest Path A suitable measure for shortest path length in a network in which the similarity of two nodes is proportional to the weight is proposed in (Antoniou & Tsompa 2008). The shortest path length from i to j, in $\mathcal{L}$ is defined as the smallest sum of the inverse weights of the links among all possible paths from i to j.

$$d_{ij} = \min_{\gamma_{ij} \in \Gamma_{ij}} \Big[\sum_{m,n \in \gamma_{ij}} \frac{1}{\mathcal{L}_{mn}} \Big]$$

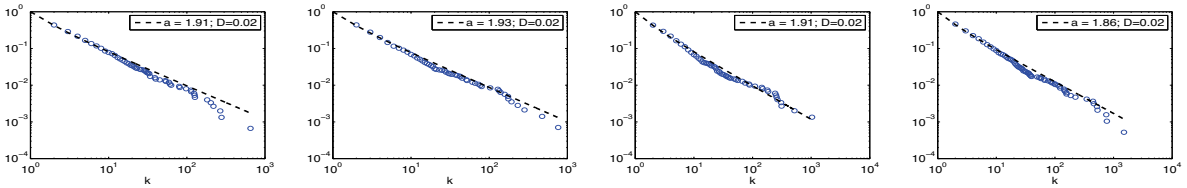
where Γ_{ij} is the set all path from i to j, and γ_{ij} denotes a single path from i to j.

Observations

In this section we will describe observations on a set of real-world as well as our method of generating perfectly homogeneous document clusters.

Cluster Name	Size	Range	Source	Regular Expression
DP	938	1965–2007	AAN	``(P|p)ars(e|ing)''
MT	844	1965–2007	AAN	``(T|t)ranslat(e|ing|ion)''
OO	1,035	1980–2008	DBLP	``(O|o)bject[-](O|o)riented ((L|l)ang|(P|p)rogram)''
DB	1,509	1975–2008	DBLP	``(R|r)elational (D|d)atabase''

Table 1: The set of title collections extracted from AAN and DBLP



$(a)DP$ $(b)MT$ $(c)OO$ $(d)DB$

Figure 2: Term frequency vs. Rank of all terms in (a)DP, (b)MT, (c)OO, and (d)DB plotted on a log-log scale. Slopes of the best fitted curves are -1.91 $(D = 0.02)$, -1.93 $(D = 0.02)$, $-1.91(D = 0.02)$, and -1.86 $(D = 0.02)$ respectively.

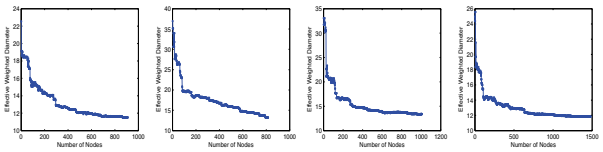
Real-world Datasets

To look at different homogeneous text collections, we use the DBLP and the ACL Anthology papers. We extracted papers in 4 different topics from the DBLP and AAN corpora. The ACL Anthology[1]http://aclweb.org/anthology-new/ is a collection of papers from the Computational Linguistics journal, as well as proceedings from ACL conferences and workshops. DBLP[2]http://www.informatik.uni-trier.de/ ley/db/ is the largest Computer Science bibliography archive. The DBLP XML records that we used to extract smaller clusters in this paper, contains $1,095,872$ articles ranging from 1937 to 2009.

Each cluster is a set of chronologically sorted titles in DBLP or AAN in which the topic phrase is matched within the title of the papers. Table 1 shows the number of articles, publication range, and the source of the 4 clusters, as well as the regular expressions used to extract each of them.

Term Frequency Distribution Figure 2 illustrates the term frequency of each index term that has the given frequency rank for our text collections on a log-log scale. These plots reveal the fact that the terms appear obeying Zipf's law (Zipf 1949) of the form $freq(k; a, V) = \frac{\frac{1}{k^a}}{\sum_{n=1}^{N} \frac{1}{n^a}}$, where V is the vocabulary size, k is the rank, and a is the exponent of the distribution, which is 1 in the basic version of Zipf's law. In these plots, D is the Kolmogorov-Smirnov goodness of fit statistic.

Densification For each of the datasets we create a latent network in which nodes are article titles and weighted edges are the cosine similarities. At each time t, the network contains an article t and all other papers published before that. For each network, $\mathcal{L}$, we study the number of nodes $v(t)$ and the sum of weights $e(t)$, at each point in time, t. The sum of weights in a latent network can be interpreted as the number of edges in a regular social network. We observe the densification power-law in latent networks with the following property: $e(t) \propto v(t)^a$



$(a)DP$ $(b)MT$ $(c)OO$ $(d)DB$

Figure 4: Effective weighted diameter versus the number of nodes in (a)DP, (b)MT, (c)OO, and (d)DB

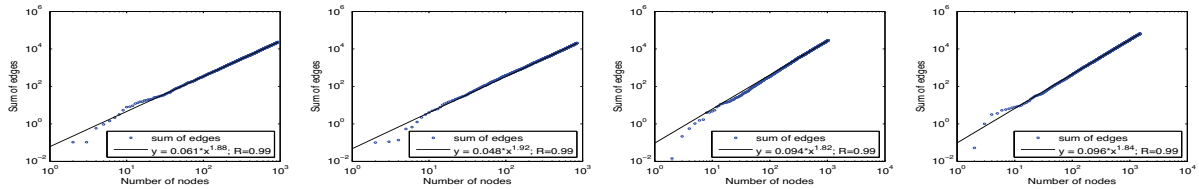
Figure 3 shows the sum of weights versus the number of nodes in each of the 4 datasets on log-log scale. These plots exhibit slopes that are all significantly greater than 1. This confirms a non-linear growth in the sum of the edge weights versus the number of nodes. In these plots, R is the correlation coefficient of between the actual values and the predicted values. Such a densification power-law in a network should result in the emergence of shrinking diameters. Shrinking diameters have been observed before in other networks (Leskovec, Kleinberg, & Faloutsos 2005; Kumar, Novak, & Tomkins 2006) but not in latent networks. Figure 4 shows the effective weighted diameter, which is the distance below which 90 percent of all shortest paths fall, in the 4 growing latent networks. These figures confirm that latent networks exhibit shrinking diameters while they grow over time.

Synthetic Homogeneous Datasets

In this section we investigate if our findings hold for a set of synthetic documents that represent a perfectly homogeneous cluster. Documents that cover smaller number of topics and are more similar to each other form a collection which we refer to, as a *homogeneous cluster*. We use certain assumptions about homogeneity in text collections, based on which we generate synthetic document collections. More precisely,

[1]http://aclweb.org/anthology-new/

[2]http://www.informatik.uni-trier.de/ ley/db/



$(a)DP$ $(b)MT$ $(c)OO$ $(d)DB$

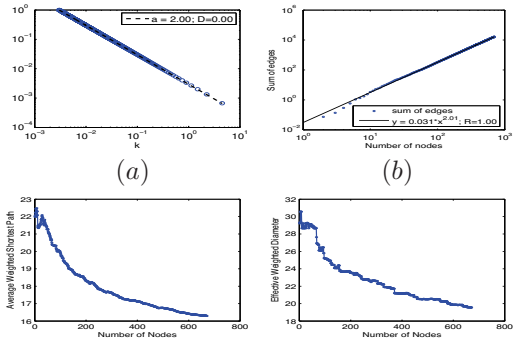
Figure 3: Latent network growth. The sum of edges versus the number of nodes in (a)DP, (b)MT, (c)OO, and (d)DB plotted on a log-log scale. Slopes are 1.88 ($R = 0.99$), 1.92 ($R = 0.99$), 1.82 ($R = 0.99$), and $R = 1.84$ ($R = 0.99$) respectively.

we assume in a homogeneous collection,

1. The number of occurrences of a term in documents of the collection is a Poisson random variable.
2. The frequency of terms in the collection follows the Zipfian distribution.

Documents are written in a way to contain information about a certain set of topics. The extent to which a document covers topics is different for different topics. Our first assumption indicates that the frequency of a given term in a perfectly homogeneous cluster is determined by a Poisson random variable. The intuition behind this assumption is based on the N-Poisson model of term frequency, introduced and evaluated by (Margulis 1992). This model argues that the frequency of the occurrence a term in a particular document of a cluster depends on the extent to which the document is related to the topic that is associated with that term. This frequency is the sum of N Poisson distributions, in which every summand is an independent single Poisson. Each single Poisson in this sum describes the frequency of the term within the subset of documents that belong to the same level of coverage of the topics related to the term.

More formally, the frequency of the term w_i in document d_j is a random variable described with the density $P(freq(w_i, d_j) = k) = \sum_l \pi_l \frac{\lambda_l^k}{k!} e^{-\lambda_l}$. Here l denotes the class of coverage of the topic regarding the term w_i, and λ_l is the average number of occurrences of the term regarding class l. In this representation π_l is the probability that the document in the collection belongs to the class l, and so $\sum_l \pi_l = 1$. The distribution of the term w_i within the class l is governed by a single Poisson process with a mean of $\lambda_{l,i}$, and thus the distribution of w_i in documents within the whole collection is governed by the sum of Poisson distributions, one for each class of coverage (Margulis 1992). It follows that, in a perfectly *homogeneous* cluster in which documents only cover one topic, the N-Poisson model reduces to a single Poisson model. In such situation, the distribution of term w_i within the whole collection of documents is governed by a single Poisson process with a unique mean, say λ_i: $P(freq(w_i, d_j) = k) = \frac{\lambda_i^k}{k!} e^{\lambda_i}$. The expected number of times that a term w_i appears in a document d_j is then equal to $E(freq(w_i, d_j)) = \lambda_i$. Let's assume that we have a constant number of documents in our collection equal to D. Then the expected total number of times that a term w_i appears in a homogeneous cluster is $D \cdot \lambda_i$, where λ_i is the



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

Figure 5: (a) Zipf's law, (b) The sum of the edges vs. the number of nodes, (c) Average weighted shortest path vs. the number of nodes, and (d) Effective weighted diameter vs. the number of nodes for the synthetic homogeneous cluster.

expected value of the number of times the term w_i appears in a document. Let's also assume that the total number of distinct terms in our collection (i.e. vocabulary size) is V.

For convention, assume that the term frequency vector is sorted in a decreasing order in which the first word is the most frequent word of the vocabulary. That is, in the collection C, according to the basic version of the Zipf's law with exponent 1, $freq(w_i, C) = \frac{freq(w_1, C)}{i}$, for $1 \leq i \leq V$, and so,

$$E(freq(w_i, C)) = E(\frac{freq(w_1, C)}{i})$$

$$\Rightarrow \quad E(\sum_j freq(w_i, d_j)) = \frac{E(\sum_j freq(w_1, d_j))}{i}$$

$$\Rightarrow \quad D \cdot \lambda_i = \frac{D \cdot \lambda_1}{i} \Rightarrow \lambda_i = \frac{\lambda_1}{i}$$

The above argument indicates that according to the Zipfian assumption and the Poisson model of occurrences, the expected number of appearances of terms is inversely proportional to their rank in the collection. This also shows that the most frequent term of the collection has the highest expected number of occurrences in each document.

Let S_C be the total number of terms in the entire collection. Then, $E(S_C)$ denotes the expected value of the number

of terms in the collection. The expected value of the average document length, $\bar{\ell}$, in the collection is then calculated as

$$E(\bar{\ell}) = E(\frac{S_C}{D}) = \frac{E(\sum_{i=1}^{V} freq(w_i, C))}{D}$$
$$= \frac{\sum_{i=1}^{V} D.\lambda_i}{D} = \sum_{i=1}^{V} \lambda_i = \sum_{i=1}^{V} \frac{\lambda_1}{i} = \lambda_1 \sum_{i=1}^{V} \frac{1}{i}$$

Solving for λ_i results in $\lambda_i = \frac{E(\bar{\ell})}{i \cdot \sum_{i=1}^{V} \frac{1}{i}}$. Based on the above argument, we can generate synthetic documents related to a single topic as well as their corresponding cosine matrix. Based on the two parameters, the average length of the documents and the size of vocabulary, we create a vector of Poisson random numbers with pre-computed means, $\lambda_i s$, to represent each document. Using the generated term frequency vectors for each document, we are able to compute the cosine similarity matrix for the synthetic collection and observe its evolution.

Figure 5 shows the rounded expected Term frequency versus rank (a), the sum of the edges vs. the number of nodes (b), average weighted shortest path vs. the number of nodes (c), and the effective weighted diameter vs. the number of nodes (d), in a synthetic homogeneous cluster of 700 documents with the average document length of 50 words chosen from a vocabulary of size $1,500$. This figure shows that a synthetic homogeneous cluster undergoes a similar growth behavior as it grows. Here, we assume that all documents in a cluster select words from a common Zipfian distribution of terms, that describes the term frequency of the entire cluster.

Related Work

Social Networks In past decade, several evolution models have been proposed for social networks and their properties (Barabási & Albert 1999; Kumar *et al.* 2000; Jin, Girvan, & Newman 2001; Leskovec, Kleinberg, & Faloutsos 2005).

Lexical Networks Various lexical networks have also been studied in several previous works (Steyvers & Tenenbaum 2005; Dorogovtsev & Mendes 2001; Menczer 2004).

Conclusion

In this work we studied latent networks built upon scientific titles over time. We showed that the new nodes attach to other nodes with similarities whose sum is not constant but grows with a power of the number of nodes which is significantly greater than 1. By looking at the effective weighted diameter, we observe that the average weighted geodesic distance decreases in the network as new nodes arrive. This means that the similarity network is becoming denser and denser over time, and that more recent titles are more similar to previous ones. We show that the observations in real-world datasets also hold in a synthetically generated and perfectly homogeneous dataset.

Acknowledgments

This paper is based upon work supported by the National Science Foundation grant iOPENER: A Flexible Framework to Support Rapid Learning in Unfamiliar Research Domains, jointly awarded to U. of Michigan and U. of Maryland as IIS 0705832. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Antoniou, I. E., and Tsompa, E. T. 2008. Statistical analysis of weighted networks. *Discrete Dynamics in Nature and Society* 2008-375452.

Barabási, A.-L., and Albert, R. 1999. Emergence of scaling in random networks. *Science* 286:509–512.

Dorogovtsev, S. N., and Mendes, J. F. F. 2001. Language as an evolving word Web. *Proceedings of the Royal Society of London B* 268(1485):2603–2606.

Erkan, G., and Radev, D. R. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)* 22:457–479.

Jin, E. M.; Girvan, M.; and Newman, M. E. 2001. Structure of growing social networks. *Phys. Rev. E* 64(4):046132.

Kumar, R.; Raghavan, P.; Rajagopalan, S.; Sivakumar, D.; Tomkins, A.; and Upfal, E. 2000. Stochastic models for the web graph. In *FOCS '00: Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, 57. Washington, DC, USA: IEEE Computer Society.

Kumar, R.; Novak, J.; and Tomkins, A. 2006. Structure and evolution of online social networks. In *ACM SIGKDD '06*, 611–617. New York, NY, USA: ACM.

Leskovec, J.; Backstrom, L.; Kumar, R.; and Tomkins, A. 2008. Microscopic evolution of social networks. In *ACM SIGKDD '05*.

Leskovec, J.; Kleinberg, J.; and Faloutsos, C. 2005. Graphs over time: Densification law, shrinking diameters and possible explanations. In *ACM SIGKDD'05*, 177–178.

Margulis, E. L. 1992. N-poisson document modelling. In *SIGIR '92*, 177–189.

Menczer, F. 2004. Evolution of document networks. *PNAS* 101(1):5261–5265.

Newman, M. E. J. 2004. Analysis of weighted networks. *Physical Review E* 70–056131.

Steyvers, M., and Tenenbaum, J. B. 2005. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science* 29(1):41–78.

Zipf, G. K. 1949. *Human Behaviour and the Principle of Least Effort*. Addison-Wesley, Cambridge, MA.