Exploiting Phase Transition in Latent Networks for Clustering

Vahed Qazvinian Dragomir R. Radev
Department of EECS
University of Michigan, Ann Arbor
{vahed,radev}@umich.edu

Abstract

In this paper, we model the pair-wise similarities of a set of documents as a weighted network with a single *cutoff* parameter. Such a network can be thought of an ensemble of unweighted graphs, each consisting of edges with weights greater than the cutoff value. We look at this network ensemble as a complex system with a temperature parameter, and refer to it as a *Latent Network*. Our experiments on a number of datasets from two different domains show that certain properties of latent networks like *clustering coefficient*, *average shortest path*, and *connected components* exhibit patterns that are significantly divergent from randomized networks. We explain that these patterns reflect the network *phase transition* as well as the existence of a community structure in document collections. Using numerical analysis, we show that we can use the aforementioned network properties to predicts the clustering Normalized Mutual Information (NMI) with high correlation ($\overline{\rho} > 0.9$). Finally we show that our clustering method significantly outperforms other baseline methods ($\overline{\text{NMI}} > 0.5$)

Introduction

Lexical networks are graphs that show relationship (e.g., semantic, similarity, dependency, etc.) between linguistic entities (e.g., words, sentences, or documents) (Ferrer i Cancho and Solé 2001). One specific type of lexical networks include those in which edges represent a similarity relation between documents. These networks are fully connected, weighted, and symmetric (if the similarity measure is symmetric).

If we apply a cutoff value $c \in [0, 1]$, and prune the edges with values smaller than c, we will have an ordinary binary lexical network (i.e., an unweighted network in which edges denote a binary relationship). Therefore, at each value c, we have a different network. In other words, binding a network with a cutoff parameter c on edge weights as the single parameter of the network, will result in an ensemble of networks with different properties. We refer to this ensemble of networks as a *latent network*. More accurately, a latent network, $\mathcal{L}$, is an ensemble of lexical networks that are originated from the same document collection and differ by the value of a single parameter.

In our work, we analyze different properties of latent networks when the cutoff value changes, and will discuss how the network undergoes different phases and exhibits high degrees of community structure. Finally, we propose a predictive model to estimate the best cutoff value for which the network community structure is maximum and use this estimation for clustering the document collection.

Data

For our experiments, we use the data from (Qazvinian and Radev 2011) on collective discourse, a collective human behavior in content generation. This data contains 50 different datasets of collective discourse from two completely different domains: news *headlines*, and scientific *citation sentences*. Each set consists of a number of unique headlines or citations about the same non-evolving news story or scientific paper.

Table 1 lists some of these datasets with the number of documents in them.

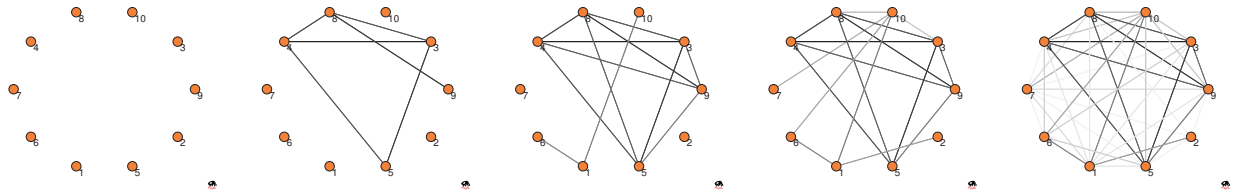
ID	type	Name	Story/Title	#
1	hdl	miss	Venezuela wins miss universe 2009	125
2	hdl	typhoon	Second typhoon hit philippines	100
3	hdl	russian	Accident at Russian hydro-plant	101
...	...	...	...	
25	hdl	yale	Yale lab tech in court	10
26	cit	N03-1017	Statistical Phrase-Based Translation	172
27	cit	P02-1006	Learning Surface Text Patterns ...	72
28	cit	P05-1012	On-line Large-Margin Training ...	71
...	...	...	...	
50	cit	H05-1047	A Semantic Approach To Recognizing TE	7

Table 1: The datasets and the number of documents in each of them (hdl = headlines; cit = citations)

Annotation

Following (Qazvinian and Radev 2008), we asked a number of annotators to read each set and extract different *facts* that are covered in each sentence. Each fact is an aspect of the news story or a contribution of the cited paper.

For example, one of the annotated datasets, `Yale`, is the set of the headlines about a murder incident at Yale. The manual annotation of the `Yale` dataset has resulted in 4 facts or classes:



$(a)\tau = 0.00$ $(b)\tau = 0.25$ $(c)\tau = 0.50$ $(d)\tau = 0.75$ $(e)\tau = 1.00$

Figure 1: Lexical network for the `Yale` dataset at 5 different τ values

ID	sentence	f_1	f_2	f_3	f_4
1	annie le slay suspect raymond clark due in court	1	1	1	0
2	attorneys to spar today over sealed annie le file	1	0	0	0
3	former yale lab tech due in court	0	1	1	0
4	former yale lab tech due in court for murder charge	0	1	1	1
5	photos: accused yale lab tech due in court today	0	1	1	0
6	raymond clark due in court	0	1	1	0
7	suspect in yale student killing to enter plea	1	0	1	1
8	yale lab tech murder suspect expected	0	1	0	1
9	yale lab tech murder suspect expected to plead not guilty	0	1	1	1
10	yale slaying suspect due in court	0	0	1	0

Table 2: Full Annotation of the `Yale` dataset results in a fact distribution matrix of sentences.

f_1 : {annie le, yale student}

f_2 : {former yale lab tech, raymond clark}

f_3 : {plea, court}

f_4 : {murder, killing}

Table 2 shows the headlines with sentence-to-fact assignments in the `Yale` dataset. The full annotation of each dataset results in a number of facts (representing classes) and a fact distribution matrix.

Network Properties

One way to look at a latent network is to use a physical point of view. The network is a complex system, and the temperature of this system will determine the interaction of the nodes. Here, nodes with smaller similarities will join each other at higher temperatures. In fact, the temperature of this system can be interpreted as

$$\tau = 1 - \text{cutoff} \tag{1}$$

increasing which will cause more nodes to connect to each other.

Figure 1 shows the cosine similarity-based latent network for the 10 documents in the `Yale` dataset at 5 different τ values. At $\tau = 0$ (cutoff = 1.00) all the edges are pruned and the network is empty, while on the other end of the spectrum, where $\tau = 1$ all edges with positive weights are present.

A simple 2-D visualization of a latent network does not reveal much information about it. Describing different aspects of the network structure is easier when looking at quantitative network properties. We observe some of the latent networks' properties over different network temperatures. Starting at $\tau = 0$ and gradually increasing it till it reaches $\tau = 1$ will cause more edges to emerge and network properties to change.

Number of Edges

Increasing the temperature τ (and thus decreasing the cutoff) will cause different edges to appear in the network according to the distribution of edge weights. To compare the number of edges in different networks we use the normalized number of edges at each τ based on Equation 2, in which $e(\tau)$ is the number of edges at temperature τ, and n is the total number of documents (nodes).

$$ne(\tau) = \frac{2e(\tau)}{n(n-1)} \tag{2}$$

Number of Connected Nodes

Another property that we are interested in is the number of nodes that have positive degrees at each τ. The number of connected nodes quantifies the distribution of $e(\tau)$ edges between n nodes. Here, we normalize this number by the total number of nodes in the graph based on Equation 3.

$$nn(\tau) = \frac{|\{i|k_i(\tau) > 0\}|}{n} \tag{3}$$

where $k_i(\tau)$ is the degree of node i at temperature τ.

Connected Components

A *connected component* (cc) of a graph is a subgraph in which there is a path between any two node pairs. The pattern in which smaller components merge into larger components or join the *largest connected component* (lcc) can quantify community structure in a network. Here, we observe the number of different connected components and the size of the largest connected component at each network temperature τ.

$$ncc(\tau) = \frac{\#\ cc(\tau)}{n}; \qquad nlcc(\tau) = \frac{|lcc(\tau)|}{n} \tag{4}$$

In a network, where community structure is weak, new nodes join the largest connected component one-by-one, and the giant component includes most of the nodes in the graph. However, in a network with an inherent community structure, we expect to see the formation of smaller separate connected components that will only merge in high temperatures.

Average Shortest Path and Diameter

In graph theory, the shortest path between two vertices is path with the smallest number of edges. In network analysis, the average shortest path (asp) of a network is the mean of all shortest path lengths between reachable vertices. Moreover, the *diameter* (d) of a network is defined as the length of the longest shortest path. We observe the normalized average shortest path ($nasp$) and the normalized diameter (nd) of each network at different values of τ.

$$nasp(\tau) = \frac{asp(\tau)}{n}; \qquad nd(\tau) = \frac{d(\tau)}{n} \tag{5}$$

Clustering Coefficient

The clustering coefficient of a graph measures the number of closed triangles in the graph. The clustering coefficient describes how likely it is that two neighbors of a vertex are connected. In social networks, it can represent the idea that "the friends of my friends are my friends." (Newman 2003a).

Watts and Strogatz's definition of clustering coefficient (Watts and Strogatz 1998) is based on a local clustering value for each vertex that is averaged over the entire network. The clustering coefficient for a given vertex i is the number of triangles connected to vertex i divided by the total possible number of triangles connected to vertex i. More formally, in a undirected graph, if $m_i(\tau)$ is the number of i's neighbors that are connected at temperature τ, and $k_i(\tau)$ is the degree of node i at τ, then the clustering coefficient of i can be defined by Equation 6.

$$c_i(\tau) = \frac{2m_i(\tau)}{k_i(\tau)(k_i(\tau) - 1)} \tag{6}$$

The global clustering coefficient of the network is defined by Equation 7. Higher global clustering coefficient values of a network would imply the existence of groups of nodes in the network that are densely connected.

$$cc(\tau) = \frac{1}{n} \sum_i c_i(\tau) \tag{7}$$

Phase Transition

Increasing τ in a latent network will cause new edges to emerge and network properties to change. We observe these changes in non-overlapping intervals of $\tau \in [0, 1]$.

The solid black lines in Figure 2 show 4 network properties for the `Yale` dataset: clustering coefficient, average shortest path, number of connected components, and the size of the largest connected component. This figure also plots the same properties for a network of the same size and edge weights, but in which edges are randomly assigned to node pairs. We can think of this randomization as a random permutation of edges that preserves the number and the weights of edges.

Figure 2 reveals a lot of information about the structure of the `Yale` latent network. When $\tau = 0$ where the network is empty the latent network and the randomized version are identical. For values of $\tau \in [0, 0.2]$, the two networks exhibit

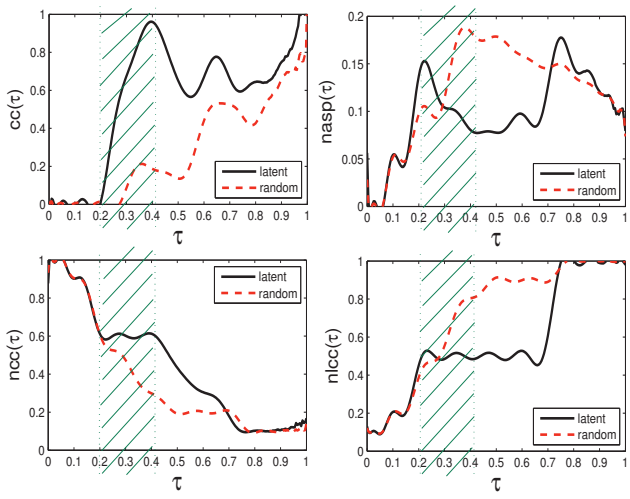


Figure 2: clustering coefficient (cc), average shortest path ($nasp$), connected components (ncc), and largest connected component ($nlcc$) in the `Yale` latent network over τ, compared with a randomized network of the same size.

similar behavior: clustering coefficient is very small, shortest path lengths increase, the number of connected components decrease and the largest connected component get bigger. However, for values of $\tau > 0.2$ the two networks show different behavior until τ is approximately greater than 0.8, where both networks become very dense and exhibit similar patterns again.

We refer to each of these intervals, in which the network has a different behavior, as a *phase*. One such phase is when the network's different connected components exhibit high degrees of community structure. The shaded area in Figure 2 ($\tau \in [0.2, 0.4]$) shows a phase in which the clustering coefficient spikes; shortest paths, unlike the randomized network, get smaller; the number of connected components is non-decreasing; and the largest connected component does not get larger. These patterns suggest the formation of dense communities in this interval because of two reasons: (1) Nodes connect to smaller components rather than the giant component. (2) Current components in the graph get denser rather than joining each other. Our goal in the rest of this paper is to predict a value $\hat{\tau}$ that best characterizes this phase, and for which the network has the best clustering of nodes represented by different connected components.

To cluster the network at each τ, we simply assign all the nodes in a connected component to the same cluster, and assign isolated (degree = 0) nodes to separate individual clusters. To evaluate this clustering we use the fact distribution matrices from the annotations and calculate the *normalized mutual information (NMI)* proposed by (Manning, Raghavan, and Schütze 2008). Let's assume $\Omega = \{\omega_1, \omega_2, \ldots, \omega_K\}$ is the set of clusters and $\mathbb{C} =$

$\{c_1, c_2, \ldots, c_J\}$ is the set of classes. Then,

$$\text{NMI}(\Omega, \mathbb{C}) = \frac{I(\Omega; \mathbb{C})}{[H(\Omega) + H(\mathbb{C})]/2} \quad (8)$$

where $I(\Omega; \mathbb{C})$ is the mutual information:

$$I(\Omega, \mathbb{C}) = \sum_k \sum_j P(\omega_k \cap c_j) \log \frac{P(\omega_k \cap c_j)}{P(\omega_k)P(c_j)} \quad (9)$$

$$= \sum_k \sum_j \frac{|\omega_k \cap c_j|}{N} \log \frac{N|\omega_k \cap c_j|}{|\omega_k||c_j|} \quad (10)$$

where $P(\omega_k)$, $P(c_j)$, and $P(\omega_k \cap c_j)$ are the probabilities of a document being in cluster ω_k, class c_j, and in the intersection of ω_k and c_j, respectively. Here, H is entropy:

$$H(\Omega) = -\sum_k P(\omega_k) \log P(\omega_k) \quad (11)$$

$$= -\sum_k \frac{|\omega_k|}{N} \log \frac{|\omega_k|}{N} \quad (12)$$

$I(\Omega; \mathbb{C})$ in Equation 9 measures the amount of information that we would lose about the classes without the cluster assignments. The normalization factor ($[H(\Omega) + H(\mathbb{C})]/2$) in Equation 8 enables us to trade off the quality of the clustering against the number of clusters, since entropy tends to increase with the number of clusters. For example, $H(\Omega)$ reaches its maximum when each document is assigned to a separate cluster. Because NMI is normalized, we can use it to compare cluster assignments with different numbers of clusters. Moreover, $[H(\Omega) + H(\mathbb{C})]/2$ is a tight upper bound for $I(\Omega; \mathbb{C})$, making NMI obtain values between 0 and 1 (Manning, Raghavan, and Schütze 2008).

The evolution of a latent network over τ can be illustrated using a dendrogram, and characterized by the quality of the clustering that the connected components produce. Figure 3 shows $\text{NMI}(\Omega, \mathbb{C})$ versus τ in the `Yale` dataset aligned with a clustering dendrogram. The shaded area in the plot ($\tau \in [0.2, 0.4]$) shows the area in which any cut on the dendrogram will result in a maximum community structure characterized by NMI.

Optimization

To find the best cut on the dendrogram, we propose a model that is similar to the Information Bottleneck method (Dai et al. 2006) in optimizing clustering mutual information.

We build an L_1-regularized log-linear model (Andrew and Gao 2007) on τ and 7 network-based features discussed before to predict $\text{NMI}(\Omega, \mathbb{C})$ at each τ. Let's suppose $\Phi : X \times Y \rightarrow \mathbb{R}^D$ is a function that maps each (x, y) to a vector of feature values. Here, the feature vector is the vector of coefficients corresponding to τ and 7 different network properties, and the parameter vector $\theta \in \mathbb{R}^D$ ($D = 8$ in our experiments) assigns a real-valued weight to each feature. This estimator chooses θ to minimize the sum of least squares and a regularization term R.

$$\hat{\theta} = arg\min_\theta \{\frac{1}{2} \sum_i ||\langle \theta, x_i \rangle - y_i||_2^2 + R(\theta)\} \quad (13)$$

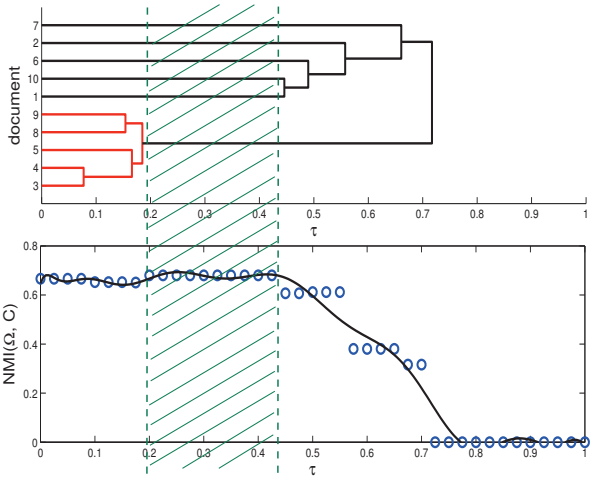


Figure 3: The dendrogram for the `Yale` dataset's latent network. Different sentences join into connected components at different temperatures.

where the regularizer term $R(\theta)$ is the weighted L_1 norm of the parameters.

$$R(\theta) = \alpha \sum_j |\theta_j| \quad (14)$$

Here, α is a parameter that controls the amount of regularization (set to 0.1 in our experiments).

To optimize the L_1-regularized objective function, we use the *orthant-wise limited-memory quasi-Newton* algorithm (OWL-QN), which is a modification of L-BFGS that allows it to effectively handle the discontinuity of the gradient (Andrew and Gao 2007). This algorithm works quite well in practice, and typically reaches convergence in even fewer iterations than standard L-BFGS (Gao et al. 2007).

NMI Prediction

Using 5-fold cross validation scheme in each category, we predict the value of $\text{NMI}(\Omega, \mathbb{C})$ ($\widehat{\text{NMI}}$) at each τ for each network. In each fold the training data consists of 20 networks. For each network $\mathcal{L}_i$, we observe 201 values of τ ($\tau \in [0, 1]$ with increments of 0.005), and calculate NMI and 7 network-based features in $\mathcal{L}_i$. We use all the observations from 20 networks to predict NMI at each τ in the test networks.

The result of this experiment is a list of $\widehat{\text{NMI}}$s at each τ for each network. Table 3 shows the average correlation between NMIs and $\widehat{\text{NMI}}$s in each category, using different features. The highest correlation is when we use all the features. However, clustering coefficient seems to play as an important indicator of the clustering quality.

Domain Adaptation

To generalize the effectiveness of network-level features in predicting cluster quality, we design the following experiment. We first use τ and 7 network features to train a model

Features	headlines		citations		Mean
$\tau + ne + nn +$	$\overline{\rho}$	95% C.I.	$\overline{\rho}$	95% C.I.	
$ncc + nlcc$	0.904	[0.844, 0.964]	0.923	[0.857, 0.989]	0.913
$nasp + nd$	0.861	[0.790, 0.932]	0.886	[0.796, 0.976]	0.873
cc	0.907	[0.856, 0.958]	0.805	[0.716, 0.894]	0.856
all	0.906	[0.845, 0.967]	0.923	[0.864, 0.982]	**0.914**
C.I. = Confidence Interval					

Table 3: Average Pearson Correlation coefficient between clustering NMI and predicted NMI at different τ values for each network, using various features

headlines		citations		Mean
$\overline{\rho}$	95% C.I.	$\overline{\rho}$	95% C.I.	
0.865	[0.786, 0.945]	0.929	[0.867, 0.991]	0.897
C.I. = Confidence Interval; Features: all.				

Table 4: Average prediction correlation when the model is trained on the other category.

on all the 25 networks from the citations category (at 201 equally spaced values of $\tau \in [0, 1]$) and use this model to predict NMI at each τ for each headline network. We also do the same experiment when the model is trained on headline networks and tested on citation networks. Table 4 reports the average correlation between predicted NMIs and actual NMIs at various τ values.

Clustering

We have shown the effectiveness of network-based features in predicting the clustering quality. Here, we employ our model to find a good clustering of a document collection. Our clustering works by simply applying the best clustering τ_c, the temperature that results in the highest predicted NMI:

$$\tau_c = \arg \max_{\tau \in [0,1]} \widehat{\text{NMI}}(\tau) \tag{15}$$

Applying τ_c to a latent network means pruning all the edges whose weight is below the cutoff from Equation 1. We then simply, assign all the nodes in each connected component to a single cluster.

Here, to build a predictive model of NMI, we follow our first experiment, and perform a 5-fold cross validation for each category. We compare the results of this experiment with 3 clustering systems: Random, Modularity-based, and K-means.

Random The Random clustering, randomly assigns each document to one of k clusters. Here we assigned k to be the number of classes in each dataset ($|f|$). Although Random is basically a weak baseline, using $|f|$ as the number of classes makes is stronger.

Modularity-based Modularity is a measure of network community division quality and is based on the measure of assortative mixing (Newman 2003b). Here we explain Newman's definition of modularity as defined in (Newman 2004; Clauset, Newman, and Moore 2004). Consider a division in the network with k communities. Let's define e as the community matrix. e is a $k \times k$ symmetric matrix in which e_{ij} is the fraction of all edges in the network that link a vertex in community i to a vertex in community j. The trace of this matrix is the fraction of edges that link vertices within the same community.

$$\text{Tr } e = \sum_i e_{ii} \tag{16}$$

A good division should result in a high value of the trace matrix. Let's also define the row sums as $a_i = \sum_j e_{ij}$, which represents the fraction of edges that connect to vertices in community i. In a random network in which edges fall between nodes regardless of any community structure, we would have $e_{ij} = a_i a_j$. In such a network, a_i^2 shows the fraction of edges within the community i. Given this setting, modularity is defined as Equation 17

$$Q = \sum_i (e_{ii} - a_i^2) \tag{17}$$

If the number of within-community edges is no better than random, we will have $Q = 0$. Higher values of Q indicates strong community structure, while $Q = 1$ is the maximum value Q can obtain.

The modularity-based algorithm (Newman 2004) uses *edge betweenness* to do the clustering. Edge betweenness in a network is an extension of the *node betweenness* definition (Freeman 1977), and measures the number of shortest paths in the graph that fall on the given edge. Intuitively, removing edges will high betweenness values will cause node pair to become more separated and form communities. Thus this algorithm iteratively removes edges with highest betweenness values, and stops when modularity is maximal.

K-means We finally used two variants of the K-means algorithm as baselines. In the first one, we run K-means on each collection with a constant number of clusters (k= 4 in our experiments), and in the second one we assign k to be the number of classes from the annotations in each dataset ($k = |f|$).

Table 5 lists the average NMI achieved by each method in each category. As this table shows the latent network model can achieve high values of NMI in clustering while outperforming other state of the art algorithms.

Method	headlines		citations		Mean
	NMI	95% C.I.	NMI	95% C.I.	
Random	0.183	[0.124, 0.243]	0.272	[0.201, 0.343]	0.227
K-means(4)	0.310	[0.244, 0.377]	0.333	[0.253, 0.413]	0.321
K-means(f)	0.364	[0.289, 0.439]	0.378	[0.298, 0.458]	0.371
Modularity	0.254	[0.193, 0.315]	0.298	[0.234, 0.362]	0.276
Latent network	**0.489**	[0.425, 0.553]	**0.575**	[0.515, 0.635]	**0.532**
C.I. = Confidence Interval					

Table 5: Average clustering Normalized Mutual Information (NMI) for each method, in each category.

Related Work

Several properties of lexical networks have been analyzed before (Ferrer i Cancho and Solé 2001; Ferrer i Cancho, Solé, and Köhler 2004). Steyvers and Tenenbaum (Steyvers

and Tenenbaum 2005) examined free association networks, WordNet, and the Roget Thesaurus, and noted five different properties in semantic networks.

The evolution of lexical networks over time has also been studied in (Dorogovtsev and Mendes 2001; Caldeira et al. 2006). These studies found that the resulting network for a text corpus exhibited small-world properties in addition to a power-law degree distribution.

It has been noted that although the standard growth models based on preferential attachment fit the degree distribution of the world wide web and citation networks, they fail to accurately model the cosine distribution of the linked documents. A mixture model for cosine distribution of linked documents is proposed in (Menczer 2004), which combines preferential attachment with cosine similarity. This model makes use of the idea that authors don't just link to the common pages on the web, but also take into account the content of these pages. Authors tend to link to and cite articles that are related to their own content. Menczer's model generates networks that reproduce the same degree distribution and content distribution of real-world information networks. They also generate networks by simulating the Open Directory Project (DMOZ) network and a collection of article published in the Proceedings of the National Academy of Sciences (PNAS). Their results show that their model not only fits the degree distribution, but it fits the similarity distribution, where the probability of a node to be linked is

$$Pr(i) = \alpha \frac{k(i)}{mt} + (1 - \alpha)\bar{Pr}(i)$$

where $i < t$ and $\alpha \in [0, 1]$ is a preferential attachment parameter.

Finally, graph based techniques have been used for other applications in NLP such as summarization (Erkan and Radev 2004), and summary evaluation (Pardo et al. 2006).

Conclusion

In this work, we define latent network, an ensemble of similarity networks between documents, and show how we can exploit its properties to predict the best cutoff at which the community structure in the network, and thus the clustering quality is maximum. We will pursue 3 ideas in future: (1) Apply the clustering technique to other tasks like text summarization and perform an extensive extrinsic evaluation of the clustering technique. (2) Extend our datasets to even wider range of document types. (3) Examine the relation between phase transition in document collections and the underlying Zipfian distribution. Such a model would enable us to explain why some certain patterns are seen in document networks but not other social networks.

Acknowledgments

This work is supported by the National Science Foundation grant number IIS-0705832 and grant number IIS-0968489. 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 supporters.

References

Andrew, G., and Gao, J. 2007. Scalable training of l1-regularized log-linear models. In *ICML '07*, 33–40.

Caldeira, S. M.; Lobão, T. C. P.; Andrade, R.; Neme, A.; and Miranda, J. 2006. The network of concepts in written texts. *The European Physical Journal B-Condensed Matter* 49(4):523–529.

Clauset, A.; Newman, M. E. J.; and Moore, C. 2004. Finding community structure in very large networks. *Phys. Rev. E* 70(6):066111.

Dai, B. T.; Koudas, N.; Ooi, B. C.; Srivastava, D.; and Venkatasubramanian, S. 2006. Rapid identification of column heterogenity. In *Proc. IEEE Intnl. Conf. Data Mining*.

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 centrality as salience in text summarization. *JAIR* 22(1):457.

Ferrer i Cancho, R., and Solé, R. V. 2001. The small-world of human language. *Proceedings of the Royal Society of London B* 268(1482):2261–2265.

Ferrer i Cancho, R.; Solé, R. V.; and Köhler, R. 2004. Patterns in syntactic dependency networks. 69(5).

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

Gao, J.; Andrew, G.; Johnson, M.; and Toutanova, K. 2007. A comparative study of parameter estimation methods for statistical natural language processing. In *ACL '07*.

Manning, C. D.; Raghavan, P.; and Schütze, H. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

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

Newman, M. E. J. 2003a. The structure and function of complex networks. *SIAM Review* 45(2):167–256.

Newman, M. J. 2003b. Mixing patterns in networks. *Rev. E 67, 026126*.

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

Pardo, T.; Antiqueira, L.; das Graças Volpe Nunes, M.; Jr., O. N. O.; and da Fontoura Costa, L. 2006. Modeling and evaluating summaries using complex networks. In *PPROPOR '06*.

Qazvinian, V., and Radev, D. R. 2008. Scientific paper summarization using citation summary networks. In *COLING 2008*.

Qazvinian, V., and Radev, D. R. 2011. Learning from collective human behavior to introduce diversity in lexical choice. In *ACL '11*.

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.

Watts, D. J., and Strogatz, S. 1998. Collective dynamics of small-world networks. *Nature* 393:440–442.