New Journal of Physics
The open-access journal for physics

Triangular clustering in document networks

Xue-Qi Cheng[1,4], Fu-Xin Ren[1], Shi Zhou[2] and Mao-Bin Hu[3]

[1] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, People's Republic of China
[2] Department of Computer Science, University College London, Malet Place, London WC1E 6BT, UK
[3] School of Engineering Science, University of Science and Technology of China, Hefei 230026, People's Republic of China
E-mail: cxq@ict.ac.cn, renfuxin@software.ict.ac.cn, s.zhou@adastral.ucl.ac.uk and humaobin@ustc.edu.cn

New Journal of Physics **11** (2009) 033019 (10pp)
Received 15 November 2008
Published 13 March 2009
Online at http://www.njp.org/
doi:10.1088/1367-2630/11/3/033019

Abstract. Document networks have the characteristic that a document node, e.g. a webpage or an article, carries meaningful content. Properties of document networks are not only affected by topological connectivity between nodes, but are also strongly influenced by the semantic relation between the content of the nodes. We observed that document networks have a large number of triangles and a high value clustering coefficient. Also there is a strong correlation between the probability of formation of a triangle and the content similarity among the three nodes involved. We propose the degree-similarity product (DSP) model, which well reproduces these properties. The model achieves this by using a preferential attachment mechanism that favours the linkage between nodes that are both popular and similar. This work is a step forward towards a better understanding of the structure and evolution of document networks.

[4] Author to whom any correspondence should be addressed.

Contents

1. Introduction

In recent years, studying the structure, function and evolution of complex networks in society and nature has become a major research focus [1]–[4]. Examples of complex networks include the Internet, the World Wide Web, the international aviation network, social collaboration between members of a group of people, protein interactions in a cell, to name just a few. These networks exhibit a number of interesting properties, such as short average distance between a pair of nodes in comparison with a large network size [1], the clustering structure where one's friends are friends of each other, and the power-law distribution of the number of connections a node has [2].

The present paper concerns a particular type of complex networks, namely document networks, such as the Web and citation networks. Document networks have the characteristic that a document node, e.g. a webpage or an article, carries text or multimedia content. The properties of document networks are not only affected by topological connectivity between nodes, but are also strongly influenced by the semantic relation between the content of nodes. Research on document networks is relevant to a number of issues, such as Web navigation and information retrieval [5]–[7].

Menczer [8] reported that the probability of linkage between two documents increases with the similarity between their content. Based on this observation, he proposed the degree-similarity mixture (DSM) model, which successfully reproduces two important properties of document networks: the power-law connectivity distribution and the increasing linkage probability as a function of content similarity. The DSM model remains one of the most advanced models for document networks.

Recently, we reported that document networks exhibit a number of triangular clustering properties; for example, they have huge numbers of triangles and high clustering coefficients, and there is a positive relation between the probability of formation of a triangle and the content similarity among the three documents involved [9]. Menczer's DSM model focuses on the connectivity and content properties between two nodes, and it produces only about 5% of the triangles in real document networks. There are a number of topology models that can produce networks with a power-law distribution of connectivity with high clustering coefficient, such as the network model in [10, 11], which was based on the balance between different types of attachment mechanisms, i.e. cyclic closure and focal closure. This model, however, does not have the ingredient of document content in its generative mechanisms and cannot reproduce content-related properties of document networks.

In the present paper, we examine and model the triangular clustering properties of document networks. In section 2, firstly we introduce two datasets of real document networks, then we define a number of metrics to quantify connectivity and content properties, and finally we review Menczer's DSM model. In section 3, we propose our degree-similarity product (DSP) model, where a node's ability for acquiring a new link is given as a *product* function of node connectivity and content similarity between nodes. In section 4, we evaluate our DSP model against real data and show that the model reproduces not only the connectivity and content properties between two nodes, but also the triangular clustering properties involving three nodes. In section 5, we conclude the paper.

2. Triangular clustering in document networks

2.1. Two datasets

In this study, we examine the following two datasets of real document networks.

- `WT10g` data, which is a webpage network where a webpage is a node and two webpages are connected if there is a hyperlink between them. The WT10g data were proposed by the annual international Text REtrieval Conference (http://trec.nist.gov) and are distributed by the CSIRO (http://es.csiro.au/TRECWeb). The data preserve the properties of the Web and have been widely used in research on information modelling and retrieval [12, 13]. The data contain 1.7 million webpages, hyperlinks among them and the text content on each webpage. We study ten randomly sampled subsets of the WT10g data. Each subset contains 50 000 webpages with the URL domain name of *.com*. (A recent study has shown that subsets sampled from different or mixed domains exhibit similar properties [9].) The observations in the present paper are averaged over ten subsets.
- `PNAS` data, which is a citation network where an article is a node and two articles are linked if they have a citation relation. It contains 28 828 articles published by the *Proceedings of the National Academy of Sciences of the United States of America* (PNAS) from 1998 to 2007. We crawled the data on the journal's website (http://www.pnas.org) in May 2008 and used each article's title and abstract as its content.

2.2. Triangle and clustering coefficient

A triangle is the basic unit for clustering structure and network redundancy [9], [14]–[18]. Triangle-related properties have been used to quantify network transitivity [14] and characterize the structural invariance across websites [18].

The most widely studied triangle-related property is the clustering coefficient, C, which measures how tightly a node's neighbours are interconnected with each other [1, 4]. The clustering coefficient is calculated as the ratio of the number of triangles formed by a node and its neighbours to the maximal number of triangles they can have. When $C = 1$ a node and its neighbours are fully interconnected and form a clique; and when $C = 0$ the neighbours do

Table 1. Evaluation of the DSM model and the DSP model against the WT10g data and the PNAS data, respectively. Topological properties shown are the number of nodes N, the number of links L, the total number of weak triangles $\triangle$ and the average clustering coefficient $\langle C \rangle$. For each model, ten networks are generated for the WT10g data and the PNAS data, respectively, and the results are averaged.

Properties	WT10g	DSM model	DSP model
N	50 000	50 000	50 000
L	233 692	233 692	$234\,020_{\pm 1228}$
$\triangle$	1266 730	$62\,503_{\pm 187}$	$1233\,308_{\pm 18\,467}$
$\langle C \rangle$	0.153	$0.062_{\pm 0.001}$	$0.121_{\pm 0.001}$
Properties	PNAS	DSM model	DSP model
N	28 828	28 828	28 828
L	40 610	40 610	$40\,580_{\pm 215}$
$\triangle$	13 544	$868_{\pm 24}$	$13\,583_{\pm 329}$
$\langle C \rangle$	0.214	$0.021_{\pm 0.0002}$	$0.139_{\pm 0.001}$

not know each other at all. The average clustering coefficient over all nodes measures the level of clustering behaviour in a network.

Note that the triangle and clustering coefficients are not trivially related. As shown in table 1, the total number of triangles, $\triangle$, in the WT10g data is almost 100 times that in the PNAS data. The density of triangles in the WT10g data, measured by $\triangle/N$ or $\triangle/L$, is also many times larger. However, the average clustering coefficient, $\langle C \rangle$, of the WT10g data is smaller than that of the PNAS data.

2.3. Content similarity and linkage probability

For a given document network, we collect keywords present in all the documents in the network and construct a keyword vector space [19, 20]. The content of a document is then represented as a keyword vector, $\overrightarrow{X}$, which gives the frequency of each keyword's appearance in the document. The content similarity, or relevance, R, between two documents, i and j, is quantified by the cosine of their vectors:

$$R_{ij} = R_{ji} = \frac{\|\overrightarrow{X_i} \cdot \overrightarrow{X_j}\|}{\|\overrightarrow{X_i}\| \cdot \|\overrightarrow{X_j}\|}. \tag{1}$$

When $R_{ij} = 1$, the content of the two documents are highly related or similar to each other; when $R_{ij} = 0$, the two documents have very little in common. The linkage probability, $P(R)$, is the probability that two nodes with content similarity R are connected in the network. It is calculated as $P(R) = M^*(R)/M(R)$, where $M(R)$ is the total number of node pairs (connected or not) whose content similarity is R, and $M^*(R)$ is the number of such node pairs that are actually connected in the network.

Figure 1(a) shows that in document networks the linkage probability increases with the content similarity, i.e. the more similar two documents are the more likely they are to be connected. For example in the PNAS citation network, if two articles have $R = 0.5$, there is

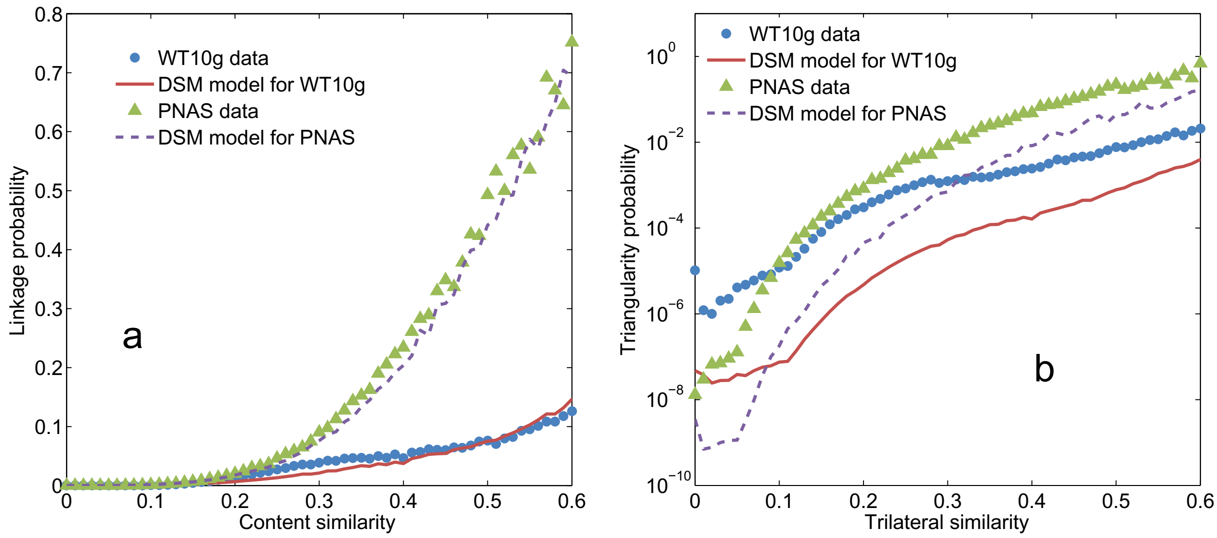

Figure 1. Linkage and triangularity probabilities for the WT10g webpage network and the PNAS citation network. The results are compared with Menczer's DSM model. (a) Linkage probability $P(R)$ as a function of content similarity R. (b) Triangularity probability $P(R^{\triangle})$, in logarithmic scale, as a function of trilateral similarity $R^{\triangle}$.

a 50% chance that they have a citation relation; by comparison, the chance is very low when $R < 0.2$.

2.4. Trilateral similarity and triangularity probability

In document networks, if a node is similar to a second node and this second node is similar to a third node, then the first and third nodes are also similar. Here, we define a new metric called the *trilateral similarity*, $R^{\triangle}$, which measures the minimum content similarity among three nodes. For three document nodes i, j and k, the trilateral similarity is the smallest (bilateral) content similarity between each pair of the three nodes, i.e.

$$R^{\triangle}_{ijk} = \min\{R_{ij}, R_{ik}, R_{jk}\}. \tag{2}$$

Similarly, we define the triangularity probability, $P(R^{\triangle})$, as the probability that three nodes with the trilateral similarity $R^{\triangle}$ form a triangle. In this study, we consider weak triangles, each of which is a circle of three nodes with at least one link (in any direction) between each pair of the three nodes.

Figure 1(b) shows that the triangularity probability is sensitive to the trilateral similarity. When the trilateral similarity $R^{\triangle}$ increases from 0.1 to 0.5, the triangularity probability increases two orders of magnitude for the WT10g data and four orders of magnitude for the PNAS data, respectively.

We note that for a given value of content similarity or trilateral similarity, the cube of the (bilateral) linkage probability provides the lower bound of the triangularity probability. But these two quantities are not trivially related because the latter is strongly determined by a network's triangular clustering structure.

Table 2. Parameters used by the two models for the datasets.

DSM model parameters	WT10g	PNAS
α	0.1	0.01
γ	3.5	3.5
DSP model parameters	WT10g	PNAS
β_1	5	7
β_2	1	4
α	10^{-12}	10^{-12}
λ	6	8

2.5. *The DSM model*

The DSM model was introduced by Menczer in 2004 [8]. The model's generative mechanism incorporates content similarity in the formation of document links. At each step, one new document is added and attached by $m = L/N$ new links to existing documents. At time step t, the probability that the new document t is attached to the existing document i is

$$Pr(i) = \alpha\frac{k_i}{mt} + (1-\alpha)\overline{Pr}(i), \qquad \overline{Pr}(i) \propto \left(\frac{1}{R_{it}} - 1\right)^{-\gamma}, \tag{3}$$

where $i < t$; k_i is the number of connections, or degree, of node i; R is calculated from the document content of the given network; γ is a constant that is calculated based on real data; and α is a preferential attachment parameter. The first term of equation (3) favours an old node, which is already well connected and the second term favours one whose content is similar to the new node. The tunable parameter $0 \leqslant \alpha \leqslant 1$ models the balance between choosing a popular node with large degree or choosing a similar node with high content similarity.

For each of the two document networks under study, we use the DSM model to grow ten networks to the same size as the real network and results are averaged over the ten networks (see table 1). Table 2 gives the model parameters that are obtained, as Menczer [8] did, by best fitting. Menczer has shown that the DSM model is able to reproduce the degree distribution of document networks. Figure 1(a) shows that the DSM model also produces a sound prediction on the relation between linkage probability and content similarity.

In terms of triangular clustering properties, table 1 shows that the model, however, produces only about 5% of the total number of triangles contained in the real networks and underestimates the average clustering coefficient of the networks. Figure 1(b) shows that the model also significantly underestimates the correlation between triangularity probability and trilateral similarity.

3. The DSP model

In the present paper, we introduce a new generative model for document networks; we call it the DSP model. Our model is partially inspired by the multi-component graph growing models of [21, 22]. The model starts from an initial seed of a pair of linked nodes. At each time step, one of the following two actions is taken:

- Growth: with probability p, a new isolated node is introduced into the network. Parameter p is a constant, which is given by the numbers of nodes and links of the generated network, $p = N/(N+L)$, and determines the average node degree of the generated network, i.e. $\langle k \rangle = 2L/N = 2(1-p)/p$.
- DSP preferential attachment: with probability $(1-p)$, a new link is attached between two nodes. The link starts from node i and ends at node j. The two nodes are chosen by the following preferential probabilities:

$$\Pi(i) = \frac{k_i^{\text{out}} + \beta_1}{\sum_m (k_m^{\text{out}} + \beta_1)}, \tag{4}$$

$$\Pi_i(j) = \frac{(k_j^{\text{in}} + \beta_2)(R_{ij}^{\lambda} + \alpha)}{\sum_l \left[(k_l^{\text{in}} + \beta_2)(R_{il}^{\lambda} + \alpha)\right]}, \tag{5}$$

where k_i^{out} is the out-degree of node i, k_j^{in} is the in-degree of node j, m and l run over all existing nodes, $l \neq i$. The content similarity R_{ij} is calculated from the document content of the given network. Parameters β_1, β_2, α and λ all take positive values. β_1 and β_2 give nodes with $k^{\text{out}} = 0$ or $k^{\text{in}} = 0$, respectively, an initial ability for acquiring links. α allows that even very different documents (with $R \simeq 0$) still have a chance to link with each other. λ tunes the weight of the content similarity in choosing a link's ending node.

It is notable that equation (5) is a product function of degree and content similarity. This ensures that links are preferentially attached between nodes that are *both* popular *and* similar. As shown in the following section, this mechanism effectively increases the chance of forming triangles among similar nodes.

4. Evaluation of the DSP model

For each of the two document networks, we generate ten networks using the DSP model with different random seeds. We avoid creating self-loops and duplicate links. The ten networks are grown to the same size as the target network. Results are then averaged over the ten networks.

As shown in table 1, the DSP model reproduces well the number of triangles and the average clustering coefficient of the two document networks. Figures 2 and 3 show that the model also closely resembles the two networks' distribution of node in-degree, linkage probability as a function of content similarity, clustering coefficient as a function of node degree, and triangularity probability as a function of trilateral similarity. The average clustering coefficient of nodes with in-degree k (see figures 3(a) and (b)) gives details of a network's triangular clustering structure.

Table 2 gives the parameters used in the modelling. The value of the parameters are tuned for best fitting. Our simulation shows that for both the real networks, the best modelling result is obtained when β_1 (in equation (4)) and β_2 (in equation (5)) take different values. This suggests that node out-degree and in-degree have different weights in choosing the starting and ending nodes of a link. The values of β_1 and β_2 for modelling the WT10g data are smaller than those for the PNAS data. This suggests that a poorly linked webpage has less difficulty in acquiring a new link in comparison with a poorly cited article. A larger value of λ is used for the PNAS data. This indicates that content similarity plays a relatively stronger role than node connectivity in the growth of the citation network.

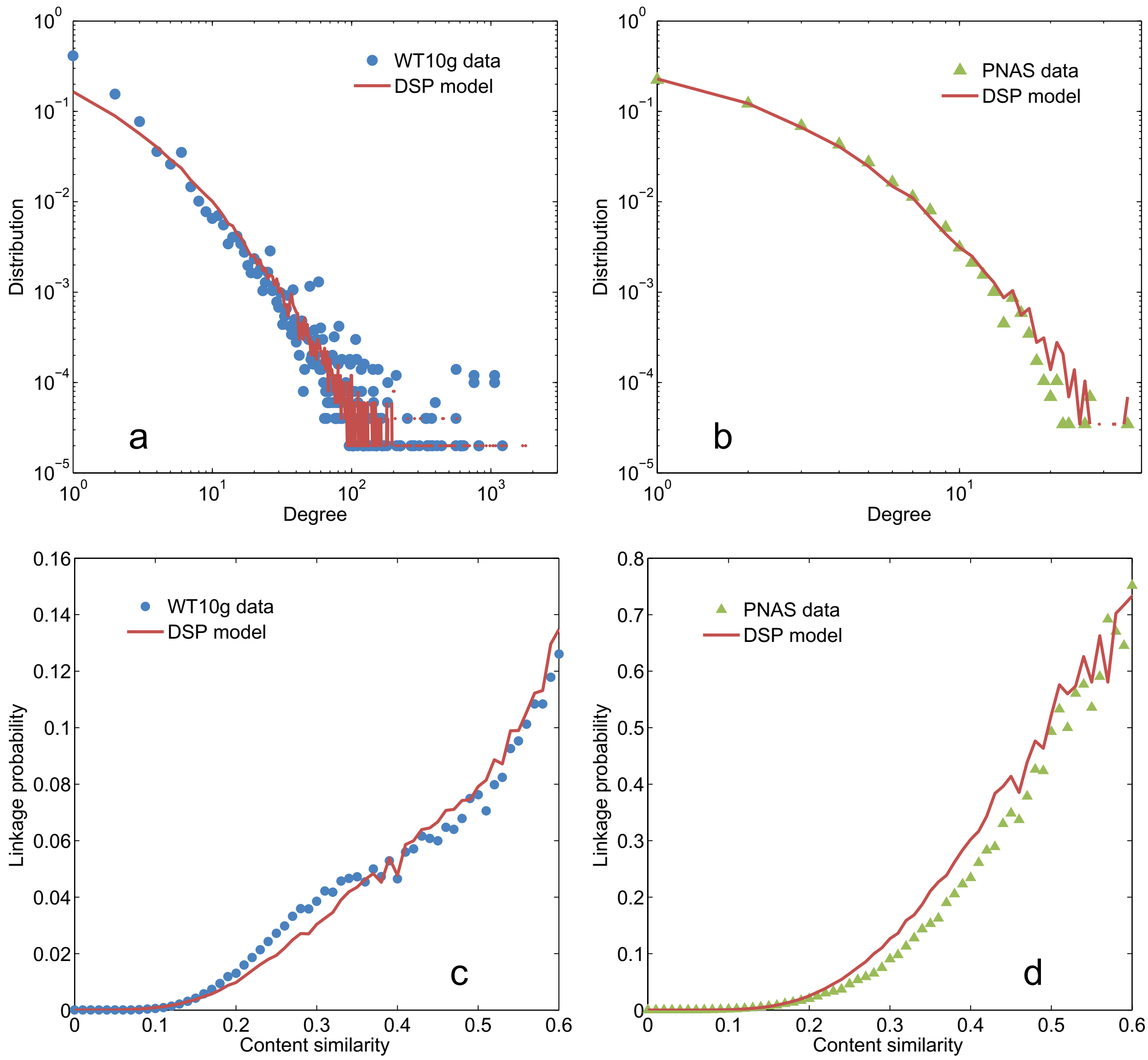

Figure 2. Evaluation of the DSP model against the WT10g webpage network and the PNAS citation network: (a, b) distribution of node in-degree; and (c, d) linkage probability as a function of (bilateral) content similarity.

5. Conclusion

It is known that document networks show a power-law degree distribution and a positive relation between the linkage probability and content similarity. In this paper, we show that document networks also contain very large numbers of triangles, high values of clustering coefficient, and a strong correlation between the triangularity probability and trilateral similarity. These three properties are not captured by the previous DSM model where a new node tends to link with an old node which is either popular or similar.

Our intuition is that a link tends to attach between two documents which are both popular and similar. We propose the DSP model that resembles this behaviour by using preferential attachment based on a product function of node connectivity and content similarity. Our model

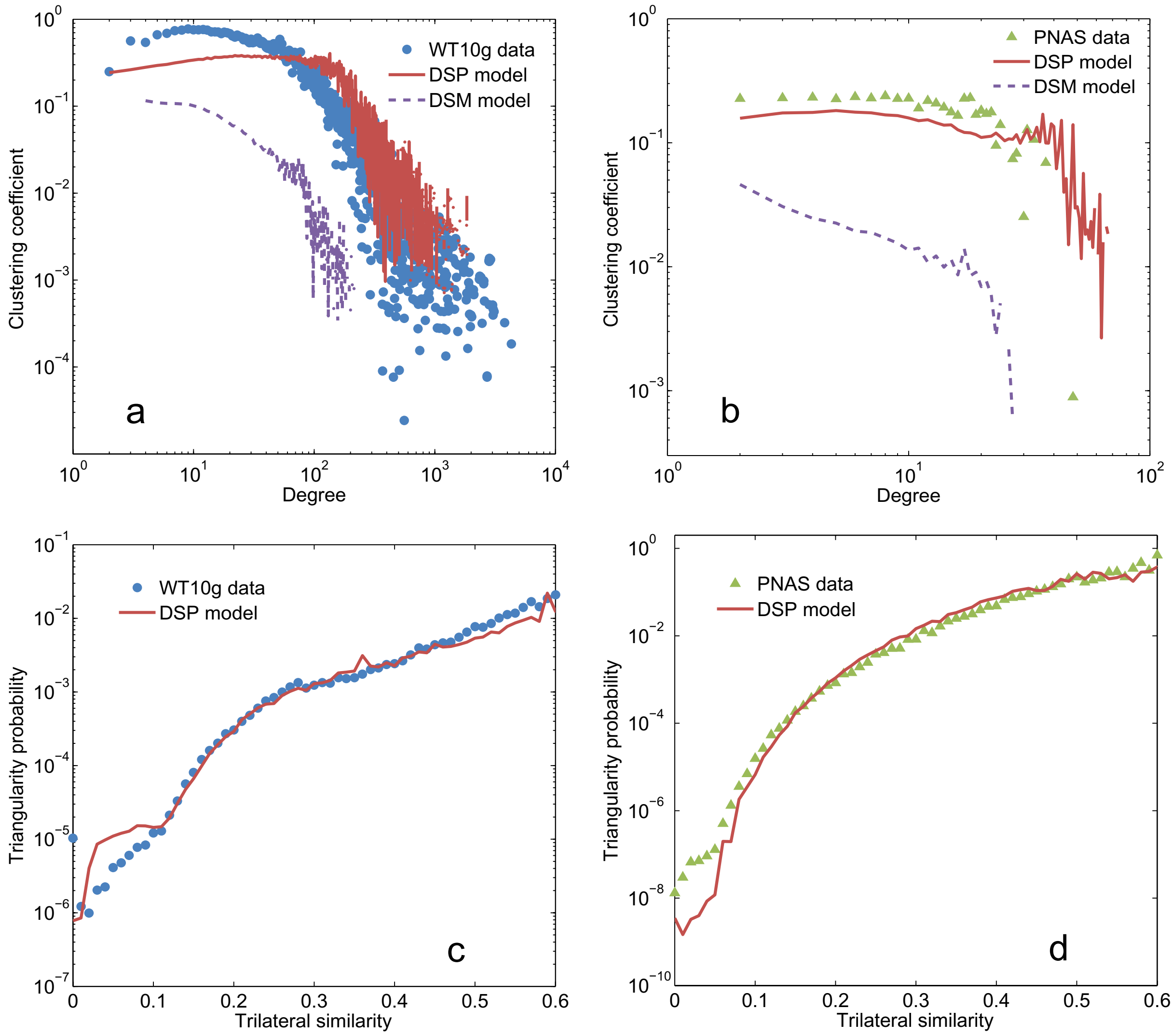

Figure 3. Evaluation of the DSP model against the two networks: (a, b) average clustering coefficient of k-degree nodes; and (c, d) triangularity probability as a function of trilateral (content) similarity.

reproduces all the above topological and content properties with remarkable accuracy. Our work provides a new insight into the structure and evolution of document networks and has the potential to facilitate research on new applications and algorithms on document networks. Future work will mathematically analyse the DSP model, examine different types of triangles in document networks and investigate the possible relation between the triangular clustering and the formation of communities in document networks.

Acknowledgments

This work was supported by the National Key Basic Research Program of China under grant number 2004CB318109 and the National Natural Science Foundation of China under grant number 60873245. SZ is supported by the Royal Academy of Engineering and the UK Engineering and Physical Sciences Research Council (EPSRC) under grant number 10216/70.

References

[1] Watts D J and Strogatz S H 1998 *Nature* **393** 440
[2] Barabási A-L and Albert R 1999 *Science* **286** 509
[3] Albert R and Barabási A-L 2002 *Rev. Mod. Phys.* **74** 47
[4] Newman M E J 2003 *SIAM Rev.* **45** 167
[5] Kleinberg J and Lawrence S 2001 *Science* **294** 1849
[6] Menczer F 2002 *Proc. Natl Acad. Sci. USA* **99** 14014
[7] Fenner T I, Levene M and Loizou G 2006 *ACM Trans. Internet Technol.* **6** 117
[8] Menczer F 2004 *Proc. Natl Acad. Sci. USA* **101** 5261
[9] Cheng X-Q, Ren F-X, Cao X-B and Ma J 2007 *Proc. IEEE/WIC/ACM Int. Conf. on Web Intelligence* p 81 http://ieeexplore.ieee.org/ielx5/4427043/4427044/04427069.pdf
[10] Toivonen R, Onnela J-P, Saramäki J, Hyvönen J and Kaski K 2006 *Physica* A **371** 851
[11] Kumpula J M, Onnela J-P, Saramäki J, Kaski K and Kertész J 2007 *Phys. Rev. Lett.* **99** 228701
[12] Soboroff I 2002 *Proc. 25th Annu. Int. ACM SIGIR Conf.* p 423 doi: 10.1145/564376.564475
[13] Chiang W-T M, Hagenbuchner M and Tsoi A C 2005 *Special Interest Tracks and Posters of the 14th Int. Conf. on the World Wide Web* p 938 doi: 10.1145/1062745.1062807
[14] Schank T and Wagner D 2005 *J. Graph Algorithms Appl.* **9** 265
[15] Serrano M A and Boguñá M 2005 *Phys. Rev.* E **74** 056114
[16] Fagiolo G 2007 *Phys. Rev.* E **76** 026107
[17] Arenas A, Fernández A, Fortunato S and Gómez S 2008 *J. Phys. A: Math. Theor.* **41** 224001
[18] Zhou S, Cox I J and Petricek V 2007 *Proc. 9th IEEE Int. Symp. on Web Site Evolution* p 73
[19] Salton G 1989 *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer* (Bostman, MA: Addison-Wesley Longman)
[20] Ricardo B-Y and Berthier R-N 1999 *Modern Information Retrieval* (Reading, MA: Addison-Wesley Longman)
[21] Callaway D S, Hopcroft J E, Kleinberg J M and Newman M E J 2001 *Phys. Rev.* E **64** 041902
[22] Krapivsky P L and Redner S 2002 *Comput. Netw.* **39** 261