Model of Complex Networks based on Citation Dynamics

Lovro Šubelj
University of Ljubljana
Faculty of Computer and Information Science
Tržaška cesta 25, SI-1000 Ljubljana, Slovenia
lovro.subelj@fri.uni-lj.si

Marko Bajec
University of Ljubljana
Faculty of Computer and Information Science
Tržaška cesta 25, SI-1000 Ljubljana, Slovenia
marko.bajec@fri.uni-lj.si

ABSTRACT

Complex networks of real-world systems are believed to be controlled by common phenomena, producing structures far from regular or random. These include scale-free degree distributions, small-world structure and assortative mixing by degree, which are also the properties captured by different random graph models proposed in the literature. However, many (non-social) real-world networks are in fact disassortative by degree. Thus, we here propose a simple evolving model that generates networks with most common properties of real-world networks including degree disassortativity. Furthermore, the model has a natural interpretation for citation networks with different practical applications.

Categories and Subject Descriptors

I.6.4 [**Computing Methodologies**]: Simulation and Modeling—*Model validation and analysis*; E.2 [**Data Structures**]: Graphs and networks

General Terms

Theory, Measurement, Experimentation

Keywords

Complex networks, graph models, degree mixing, clustering, citation networks

1. INTRODUCTION

Networks are the simplest representation of complex systems of interacting parts. Examples of these are ubiquitous in practice, including large social networks [5], information systems [18] and cooperate ownerships [22], to name just a few. Despite a seemingly plain form, real-world networks reveal characteristic structural properties that are absent from regular or random systems [23, 2]. Thus, networked systems are believed to be controlled by common phenomena.

Scale-free degree distributions [2], small-world phenomena [23], degree mixing [14] (i.e., degree correlations at links' ends) and existence of communities [6] (i.e., densely linked groups of nodes) are perhaps among most widely analyzed properties of large real-world networks. Note that community structure implies assortative (i.e, positively correlated) mixing by degree [16], which can be seen as a tendency of hubs (i.e., highly linked nodes) to cluster together. The above are also the properties captured by many random graph models proposed in the literature [9, 11, 13, 24].

However, most (non-social) networks deviate from this figure. Biological and technological networks are in fact degree disassortative (i.e., negatively correlated), while different information networks often reveal no clear degree mixing [14, 7] (see Figure 1). Thus, we here propose an evolving random graph model based on the link copying mechanism [9]. Each newly added node explores the network using the burning process in [11], while links of the visited nodes are copied independently of the latter. The model generates scale-free small-world networks with community structure and also degree disassortativity. Furthermore, it has a natural interpretation for citation networks. The above process imitates an author of a paper including references into the bibliography (i.e., its citation dynamics), which enables different practical applications in bibliometrics (see Section 3.2).

The rest of the paper is structured as follows. Section 2 introduces the proposed (Citation) model, while a thorough analysis is given in Section 3. Section 4 concludes the paper.

2. THE CITATION MODEL

Let a network be represented by a simple graph $G(N, L)$, where N is the set of nodes, $|N| = n$, and L is the set of links, $|L| = m$. Next, let Γ_i be the set of neighbors of node $i \in N$ and let k_i be its degree, $k_i = |\Gamma_i|$. Last, let k be the mean degree and k_N the mean neighbor degree.

Proposed graph model is based on the burning process of Forest Fire model [11], which we introduce first. Due to simplicity, the model is presented for undirected networks.

Let p be the burning probability, $p \in [0, \frac{1}{2})$ (see below). Initially, the network consists of a single node, while for each newly added node i, the burning process proceeds as follows.

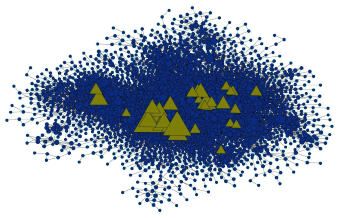
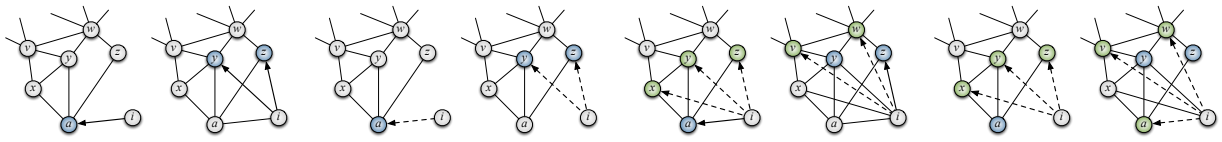


Figure 1: Data mining part of *Cora* citation network [12] with highlighted hubs (i.e., 1% of most highly linked nodes) that are scattered across the network. (Node sizes are proportional to degrees.)

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.



(a) Forest Fire model (b) Butterfly model (c) Copying model (d) Citation model (our)

Figure 2: Schematic representation of linking dynamics of different graph models. (a) In Forest Fire model [11], newly added node i selects an ambassador a (blue node) uniformly at random and links to it (solid arrow). Next, some of its neighbors are taken as the ambassadors (e.g., y and z) and the process repeats. (b) Butterfly model [13] forms links only with some fixed probability (dashed arrows). (c) In Copying model [9], node i links to a and also to some of its neighbors x, y, z (green nodes). (d) Proposed Citation model forms links only with the neighbors of the ambassador a (e.g., x and y), however, i can still link to a.

(1) i chooses an ambassador $a \in N$ uniformly at random (we say that i burns a) and links to it.

(2) i randomly selects (at most) x_p neighbors of a that were not yet burned $a_1, \ldots, a_{x_p} \in \Gamma_a$ and links to them. (x_p is sampled from a geometric distribution with mean $\frac{p}{1-p}$.)

(3) $a_1, \ldots, a_{x_p}$ are taken as the ambassadors of i (step (2)).

Since each node can be visited at most once, the burning process surely converges. Thus, to generate a network with n nodes, the model repeats the above procedure $n-1$ times.

Forest Fire model produces shrinking diameters and densification phenomena observed in temporal networks [11]. Furthermore, generated networks are scale-free and small-world, and reveal a pronounced community structure. However, in contrast to many real-world networks, the model gives degree assortative networks (see Section 3).

The model also has a natural interpretation for citation networks. Burning process imitates an author of a paper including references into the bibliography (i.e., citation dynamics). Author first reads a related paper, or selects the paper that triggered the research, and cites it (step (1)). Author then considers its bibliography for other related papers (step (2)). Some of these are further considered and also cited, while the author continues as before (step (3)). Nevertheless, Forest Fire model fails to reproduce some of the properties of citation networks (e.g., degree mixing).

Note that the described process assumes that authors read, or at least consider, all the papers they cite. However, this is indeed not the case [17]. For example, seminal work on random graphs conducted by Erdös and Rényi [4] is perhaps among most widely cited papers in network science literature. Although, presumably, only a smaller number of authors have actually read the original paper. As the work is widely discussed elsewhere, most authors have just copied the reference from another paper. On the other hand, authors also do not cite all the papers they read, though related to their work. This can be simply due to space limitations. Nevertheless, a paper can still be read thoroughly, with many of its references further considered and cited.

Examples suggest that the papers that authors read or cite are selected due to two, not necessarily dependent, processes. We thus propose a Citation model that adopts the above burning procedure to traverse the network, while the links are formed according to another independent process.

Let q be the linking probability, $q \in [0, 1)$ (see below). Initially, the network consists of a single link, while for each newly added node i, the model proceeds as follows.

(1) i chooses an ambassador $a \in N$ uniformly at random.

(2) i randomly selects (at most) x_p neighbors of a that were not yet burned $a_1, \ldots, a_{x_p} \in \Gamma_a$.

(3) i randomly selects (at most) x_q neighbors of a that were not yet linked $j_1, \ldots, j_{x_q} \in \Gamma_a$ and links to them.

(4) $a_1, \ldots, a_{x_p}$ are taken as the ambassadors of i (step (2)).

Details are the same as before. Again, the process surely converges, while the entire procedure is repeated $n-2$ times.

Let s be the mean number of burned nodes (i.e., ambassadors). A node selects $\frac{p}{1-p}$ ambassadors on each step, thus,

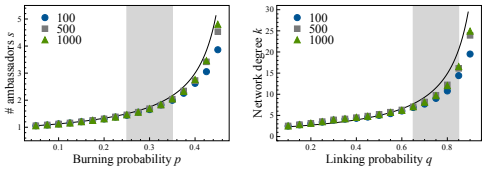
$$s \leq \sum_{x=0}^{\infty} \left(\frac{p}{1-p} \right)^x \leq \frac{1-p}{1-2p}. \tag{1}$$

A node will fail to form any link (i.e., become isolated) with probability $(1-q)^s$. Although isolated nodes are a common property of real-world networks, they are often ignored in practice or the network is even reduced to the largest connected component. Thus, for the analysis here, we repeat the procedure until the largest component has n nodes.

Since a node forms $\frac{q}{1-q}$ links on each step, expected network degree is (with $1-(1-q)^s$ correction for isolated nodes)

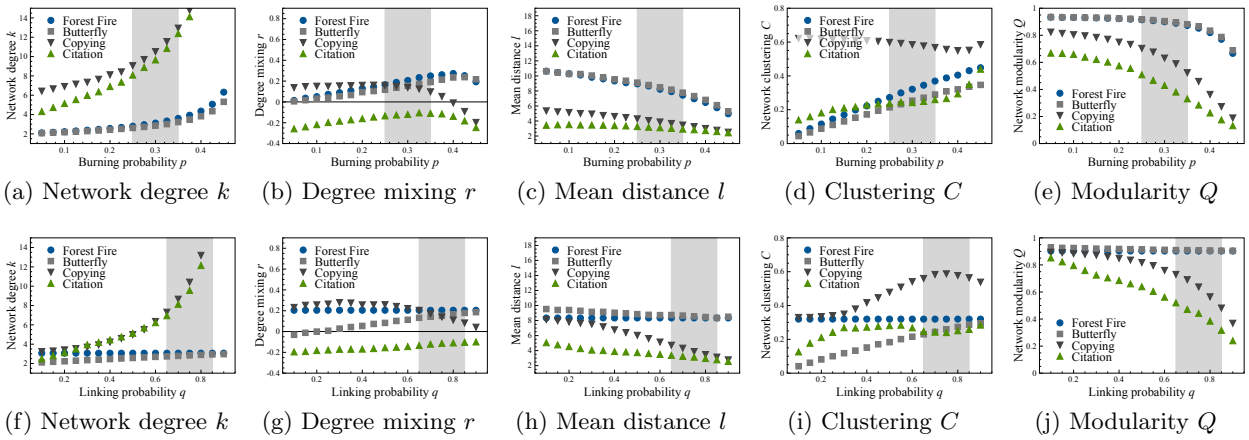
$$k \leq \frac{2qs}{1-q-(1-q)^{s+1}}. \tag{2}$$

Although equations (1) and (2) are only valid in the limit of large network size, the bounds are rather tight for large enough n (see Figure 3). Thus, given network degree k and



(a) Ambassadors s (b) Network degree k

Figure 3: Analysis of Citation model at different p and $q = 0.75$ (left), and $p = 0.3$ and different q (right). Solid lines show theoretical bounds in equations (1) and (2). (Results are estimates of the mean over 100 network realizations with different n. Shaded regions correspond to likely parameter values [10].)



(a) Network degree k (b) Degree mixing r (c) Mean distance l (d) Clustering C (e) Modularity Q

(f) Network degree k (g) Degree mixing r (h) Mean distance l (i) Clustering C (j) Modularity Q

Figure 4: Comparison of graph models at different p and $q = 0.75$ (top), and $p = 0.3$ and different q (bottom). (Results are estimates of the mean over 100 network realizations with $n = 1000$. See also caption of Figure 3.)

fixed q, one can solve the system for p, which can be used for parameter estimation in practice (see Section 3.2).

Citation model generates small-world networks with scale-free degree distribution and community structure (see Section 3.1). Furthermore, in contrast to Forest Fire model, resulting networks are degree disassortative. We stress that the key factor here is that newly added nodes do not (necessarily) link to their ambassadors, which in fact produces degree assortativity. Since a node copies the links of its ambassadors, linking to them obviously promotes assortativity. However, in the absence of an explicit process introducing assortativity, (scale-free) networks are expected to be degree disassortative [8]. The analysis in Section 3 thus also includes a variant of Forest Fire model denoted Butterfly model, where a node links to its ambassadors only with probability q (considered in [13]), as well as a variant of the proposed Citation model denoted Copying model, where a node links to each ambassador [9] (for details see Figure 2).

Other authors have proposed models very similar to ours [21, 9, 11, 13, 24]. Nevertheless, these either do not adopt the burning process to traverse the network or the model necessarily links the nodes to their ambassadors, which results in degree assortativity. More precisely, the set of the linked nodes is always a subset of the nodes burned (or vice versa). However, in the case of Citation model, these two sets can intersect arbitrarily, while they can also be disjoint.

3. EXPERIMENTAL ANALYSIS

Section 3.1 conducts an empirical analysis of Citation model and several alternatives proposed in the literature (see Section 2). Next, networks constructed with different models are compared against a larger citation network (Section 3.2).

3.1 Analysis of the model

Figure 4 shows basic statistics of the networks generated with different graph models for parameters p and q shown (see Section 2). Most notably, only the proposed Citation model gives degree disassortative networks measured by the mixing coefficient $r \in [-1, 1]$ [14] (see Figures 4(b) and 4(g)). r is simply a Pearson correlation coefficient of degrees at links' ends. Thus, $r \ll 0$ for Citation model, while $r \gg 0$ for Forest Fire and Butterfly models. Observe that Copying model also generates networks with $r < 0$ for very large p and q, however, these are much denser than comparable real-world networks (see Figures 4(a) and 4(f)).

On the other hand, all models give small-world networks with short mean distance between the nodes l [1] (see Figures 4(c) and 4(h)) and high transitivity measured by the clustering coefficient $C \in [0, 1]$ [23] (see Figures 4(d) and 4(i)). Note that C increases with p, while q has little effect on C. Furthermore, all models generate networks with clear community structure according to modularity $Q \in [0, 1]$ [15], where Q is estimated using a fast multi-stage optimization [3] (see Figures 4(e) and 4(j)). Although Q decreases with increasing p or q in the case of Citation and Copying models, the values are somewhat comparable to those observed in real-world networks. Forest Fire and Butterfly models, however, appear to overestimate Q for selected p and q.

Networks constructed with Citation model also reveal scale-free degree distributions [2] (see Figure 5(a)), thus, the model generates most common properties of real-world networks.

3.2 Cora citation network

Due to a natural interpretation for citation networks (see Section 2), the proposed model has different practical applications in bibliometrics. We here analyze author citation dynamics based on the famous *Cora* dataset [12] that contains computer science papers collected from the web, and also the references automatically parsed from the bibliographies of the papers. We extract a citation network with $n = 23166$, while other statistics are reported in Table 1.

Table 1 also includes the networks generated with Citation and Forest Fire models, where parameters p and q were es-

Table 1: Comparison of *Cora* citation network and those constructed with different graph models for p and q shown. (Results are estimates of the mean over 100 network realizations with $n = 23166$.)

Model	p	q	m	k	r
Forest Fire	0.462	-	88828	7.669	0.211
Citation	0.369	0.593	89888	7.760	−0.047
Cora			*89157*	*7.697*	*−0.055*

timated as described in Section 2. Note that Citation model well matches the disassortative mixing regime in *Cora* citation network (observe also a similar trend in Figure 5(b)), while Forest Fire model gives degree assortative networks. (For comparison based on other network properties see [19].)

Recall that s in equation (1) can be seen as the number references actually read by an author of some paper. Thus, the fraction of papers considered by the authors, relative to the number of all papers cited, can be estimated to $2s/k = 0.66$. The value is much larger than expected [17], however, the results are largely influenced by an automatic sampling procedure [12] (i.e., on average, only $k/2 = 3.85$ references of each paper are also included in the network).

4. CONCLUSION

The paper proposes a simple graph model that generates networks with most common properties of real-world networks and, in contrast to many other models, dissassortative degree mixing. The model also has a natural interpretation for citation networks with different practical applications.

Due to simplicity, the analysis in the paper is based on undirected networks. However, this presents a serious limitation, especially for citation networks considered here. Future work will extend the analysis to directed and also other types of networks, while more reliable datasets will be used for the analysis of author citation dynamics (based on *DBLP* and *WoS* data). Furthermore, the model will be rigorously compared against others with similar characteristics [20].

5. ACKNOWLEDGMENTS.

The work is supported by the Slovene Research Agency ARRS within Research Program No. P2-0359.

6. REFERENCES

[1] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. *Network flows: Theory, algorithms, and applications*. Prentice-Hall, Upper Saddle River, NJ, 1993.

[2] A. L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[3] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *J. Stat. Mech.*, P10008, 2008.

[4] P. Erdős and A. Rényi. On random graphs i. *Publ. Math. Debrecen*, 6:290–297, 1959.

[5] G. Facchetti, G. Iacono, and C. Altafini. Computing global structural balance in large-scale signed social networks. *P. Natl. Acad. Sci. USA*, 108(52):20953–20958, 2011.

[6] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *P. Natl. Acad. Sci. USA*, 99(12):7821–7826, 2002.

[7] D. Hao and C. Li. The dichotomy in degree correlation of biological networks. *PLoS ONE*, 6(12):e28322, 2011.

[8] S. Johnson, J. J. Torres, J. Marro, and M. A. Muñoz. Entropic origin of disassortativity in complex networks. *Phys. Rev. Lett.*, 104(10):108702, 2010.

[9] P. L. Krapivsky and S. Redner. Network growth by copying. *Phys. Rev. E*, 71(3):036118, 2005.

[10] P. J. Laurienti, K. E. Joyce, Q. K. Telesford, J. H. Burdette, and S. Hayasaka. Universal fractal scaling of self-organized networks. *Physica A*, 390(20):3608–3613, 2011.

[11] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data*, 1(1):1–41, 2007.

[12] A. K. McCallum, K. Nigam, J. Rennie, and K. Seymore. Automating the construction of internet portals with machine learning. *Inf. Retr.*, 3(2):127–163, 2000.

[13] M. McGlohon, L. Akoglu, and C. Faloutsos. Weighted graphs and disconnected components: Patterns and a generator. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 524–532, New York, NY, USA, 2008.

[14] M. E. J. Newman. Assortative mixing in networks. *Phys. Rev. Lett.*, 89(20):208701, 2002.

[15] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2):026113, 2004.

[16] M. E. J. Newman and J. Park. Why social networks are different from other types of networks. *Phys. Rev. E*, 68(3):036122, 2003.

[17] M. V. Simkin and V. P. Roychowdhury. Read before you cite! *Compl. Syst.*, 14:269–274, 2003.

[18] L. Šubelj and M. Bajec. Community structure of complex software systems: Analysis and applications. *Physica A*, 390(16):2968–2975, 2011.

[19] L. Šubelj and M. Bajec. Clustering assortativity, communities and functional modules in real-world networks. *e-print arXiv:12082518v1*, pages 1–21, 2012.

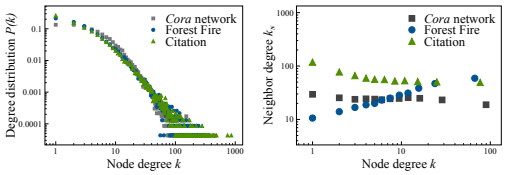
[20] L. Tan, J. Zhang, and L. Jiang. An evolving model of undirected networks based on microscopic biological interaction systems. *J. Biol. Phys.*, 35(2):197–207, 2009.

[21] A. Vázquez. Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Phys. Rev. E*, 67(5):056104, 2003.

[22] S. Vitali, J. B. Glattfelder, and S. Battiston. The network of global corporate control. *PLoS ONE*, 6(10):e25995, 2011.

[23] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[24] Z.-X. Wu and P. Holme. Modeling scientific-citation patterns and other triangle-rich acyclic networks. *Phys. Rev. E*, 80(3):037101, 2009.



(a) Distribution $P(k)$ (b) Neighbor degree k_N

Figure 5: Comparison of *Cora* network and those constructed with different graphs models (Table 1). Exponent α for a power-law fit $P(k) \sim k^{-\alpha}$ equals 3.3 for *Cora* network and 2.5 for graph models. (Nodes in (b) are aggregated into equally-sized bins.)