Journal of Complex Networks (2017) **5**, 581–603
doi: 10.1093/comnet/cnx001
Advance Access Publication on 26 March 2017

Measuring and modeling bipartite graphs with community structure

SINAN G. AKSOY
Department of Mathematics, University of California, San Diego, CA 92093, USA

TAMARA G. KOLDA[†]
Sandia National Laboratories, Livermore, CA 94550, USA
[†]Corresponding author. Email: tgkolda@sandia.gov

AND

ALI PINAR
Sandia National Laboratories, Livermore, CA 94550, USA

Edited by: Aaron Clauset

[Received on 7 July 2016; editorial decision on 12 January 2017; accepted on 13 February 2017]

Network science is a powerful tool for analyzing complex systems in fields ranging from sociology to engineering to biology. This article is focused on generative models of large-scale *bipartite graphs*, also known as two-way graphs or two-mode networks. We propose two generative models that can be easily tuned to reproduce the characteristics of real-world networks, not just qualitatively but quantitatively. The characteristics we consider are the degree distributions and the metamorphosis coefficient. The metamorphosis coefficient, a bipartite analogue of the clustering coefficient, is the proportion of length-three paths that participate in length-four cycles. Having a high metamorphosis coefficient is a necessary condition for close-knit community structure. We define edge, node and degreewise metamorphosis coefficients, enabling a more detailed understanding of the bipartite connectivity that is not explained by degree distribution alone. Our first model, bipartite Chung–Lu, is able to reproduce real-world degree distributions, and our second model, bipartite block two-level Erdös–Rényi, reproduces both the degree distributions as well as the degreewise metamorphosis coefficients. We demonstrate the effectiveness of these models on several real-world data sets.

Keywords: bipartite generative graph model; two-way graph model; two-mode network; metamorphosis coefficient; bipartite clustering coefficient; complex networks.

1. Introduction

Network science is a powerful tool for analysing complex systems in fields ranging from sociology to engineering to biology. The ability to develop realistic models of the networks is needed for several reasons. Pragmatically, we need to generate artificial networks to facilitate sharing of realistic network data while respecting concerns about privacy and security of data. More generally, generative models enable unlimited network data generation for computational analysis, for example, varying the characteristics of the graph to test a graph algorithm under different scenarios. Ultimately, we hope to *understand* the underlying nature of complex systems, and modelling them mathematically is a way to test our understanding.

This article is focused on generative models of *bipartite graphs*, also known as *two-way graphs* or *two-mode networks*. Many real-world systems are naturally expressed as bipartite graphs. The defining

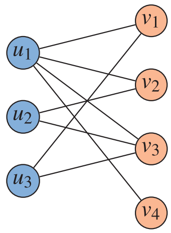


FIG. 1. Bipartite graph.

characteristic of a bipartite graph is that its vertices are divided into two partitions, U and V, such that edges, E, only connect vertices across the two partitions, that is

$$G = (U, V, E) \quad \text{with} \quad U \cap V = \emptyset \quad \text{and} \quad E \subseteq U \otimes V.$$

Examples of bipartite graphs include author–paper networks, user–product purchase histories, user–song play lists, actor–movie connections, document–keyword mappings and so on. Hypergraphs can be represented as bipartite graphs in the sense of an *incidence graph*: the nodes and hyperedges are represented by U and V, respectively, and edge (i,j) exists if node i is in hyperedge j. Bipartite graphs have been widely studied; see [1–6] and references therein. An example bipartite graph is shown in Fig. 1.

We propose a generative model that can be easily tuned to reproduce the characteristics of real-world networks, not just qualitatively but quantitatively. The measurements we consider in this article are the degree distributions and the bipartite analogue of the clustering coefficient. Of course, there are many other measurements that we could consider, and we come back to those shortly. The degree distribution is the number (or proportion) of nodes of degree d for each $d = 1, 2, \ldots$. A bipartite graph has *two* degree distributions, one for each vertex partition. As we see in the results in Section 4.3, these two distributions may be quite different from one another, in part because the size (and, consequently, the average degree) for each partition may be quite different. The clustering coefficient of a one-way graph, introduced by Holland and Leinhardt [7], is the probability that a two-path (or *wedge*) participates in a three cycle (or *triangle*); i.e.,

$$c = \frac{3n^{\triangle}}{n^{\wedge}} = \frac{3 \times (\text{total number of triangles})}{\text{total number of wedges}}.$$

One characteristic of a bipartite graph is that is has no odd-length cycles; hence, it cannot have a triangle. Robins and Alexander [5] propose a bipartite clustering coefficient that is the probability that a bipartite three-path (or *caterpillar*) participates in a bipartite four cycle (or *butterfly*), i.e.,

$$c = \frac{4n^{\text{⧖}}}{n^{\Sigma}} = \frac{4 \times (\text{total number of butterflies})}{\text{total number of caterpillars}}. \tag{1.1}$$

We call c in (1.1) the *metamorphosis coefficient* in reference to the notion of how often caterpillars become butterflies. Opsahl [8] says “it could be considered a measure of reinforcement between two individuals rather than clustering of a group of individuals.” The multiplier of four in the numerator is because every butterfly contains four distinct caterpillars, just as every triangle contains three distinct wedges; see Fig. 2.

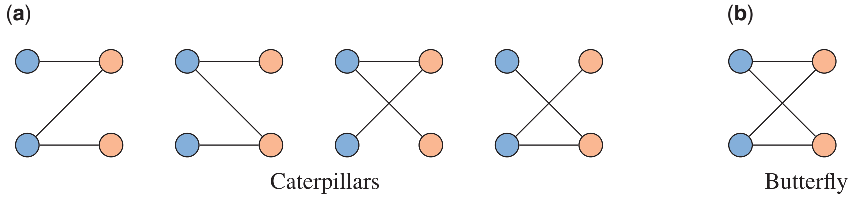


FIG. 2. A butterfly contains four distinct caterpillars.

A high metamorphosis coefficient in a bipartite graph is indicative of greater community structure in the graph, analogous to the role of a high clustering coefficient in a one-way graph. The metamorphosis coefficient plays an important role in communities because any tight-knit community will have a high metamorphosis coefficient. In fact, Sariyuce and Pinar [9] have recently shown that butterflies are the key to identifying dense substructures, corresponding to highly connected communities, in bipartite graphs. In Section 5.1, we extend (1.1) to define edge, node and degreewise metamorphosis coefficients, enabling a more detailed understanding of the bipartite community structure.

As mentioned above, there are numerous more complex metrics that can be used in the analysis of bipartite graphs; see, e.g., [4, 5]. Examples of such metrics include graph diameter (i.e., the maximum distance between any pair of nodes), singular values of the adjacency matrix, centrality measures, joint degree distributions (i.e., the proportion of nodes of degree d that connect to nodes of degree d' for all possible pairs (d, d')), assortativity (i.e., degree correlations), subgraph census (i.e., counts of all patterns of three nodes, four nodes, and so on). Barber [10] introduced a *bipartite* modularity coefficient, extending one-way Newman–Girvan modularity graph modularity to the bipartite case. Using this measure requires a partitioning of the vertices into communities, and finding a partition that maximizes Barber's modularity coefficient has been shown to be NP-hard [11]. Other metrics result by considering the *product graph* that is produced by considering just *one* vertex partition and connecting any two nodes that share a common neighbour in the bipartite graph. However, any analysis of bipartite community structure based solely on product graphs is fundamentally limited due to the structural information lost in the product graph [3, 4, 9, 12]. Because these quantities are often expensive to compute for large graphs, we do not consider them in this article. Instead, we consider the simpler metrics of degree distribution and metamorphosis coefficients here for reasons analogous to those considered by [4, 13]. The degree distributions are interesting because many large real-world graphs (including the ones we consider here) do not have the Poisson distributions that occur if edges are distributed at random; instead, they have heavy-tailed distributions (hypothesized to be power law [14] or log-normal [15]). The metamorphosis coefficient is interesting because random graphs that have the same degree distribution as a real-world graph tend to have a much lower metamorphosis coefficients, as we show in Section 4.4. This is a well-known phenomenon for the clustering coefficient of one-way graphs [16]. We contend that matching metamorphosis coefficients is critical for capturing social or, more generally, interconnectedness behaviour in networks.

With the goals of capturing both degree distribution and our newly defined degreewise metamorphosis coefficients, we develop two different models. The first is a straightforward extension of Chung–Lu (CL) [17–19], which is very closely related to the configuration model. Our experimental results show that this model is effective at reproducing the degree distributions. However, the bipartite CL graphs do not produce the same metamorphosis coefficients as observed in real-world networks. Therefore, the second model we propose is a bipartite extension of the Block Two-Level Erdős–Rényi (BTER) [13, 15] model. The BTER model is a good starting point because it reproduces both the degree distribution and degreewise clustering coefficients of a given network. To do so, it groups nodes into Erdős–Rényi (ER) subgraphs, called affinity blocks, that are highly connected and so produce high-clustering coefficients. We propose a bipartite BTER that reproduces both the degree distributions and the degreewise metamorphosis coefficients. This extension is not straightforward since the affinity block concept does not carry over easily to the bipartite case, so we develop a new method for creating the blocks. Computational results for bipartite BTER show that it achieves our goals of matching both the degree distributions and the degreewise metamorphosis coefficients.

For some of the other expensive metrics mentioned above, we briefly speculate on the performance of bipartite BTER. For instance, the largest singular values should be on the same order of magnitude, since these values rely in part on degree distribution. We may expect good results in terms of the subgraph consensus, since the affinity blocks will produce other structures beyond butterflies. Conversely, the simple model has no mechanism to match the joint degree distribution or assortativity. Community structure is more complex: bipartite BTER creates butterflies that are key to small tight-knit communities, but the model as described in this article, does not necessarily produce a hierarchy of dense structures (i.e., larger yet sparse communities that agglomerate the smaller dense blocks). In particular, we stress that it produces a large number (hundreds to millions) of small affinity blocks, in contrast to methods that focus on modularity with respect to a small number (two to twenty) of communities. In Section 6, we discuss the possibility of extending the bipartite BTER model to group affinity blocks to produce a hierarchy of loosely knit communities, but we do not pursue that avenue in this work.

Before we continue, we briefly survey other models which fall into two classes. The first class is bipartite configuration models which, as mentioned above, are very similar to our proposed bipartite CL model. These can be implemented efficiently and scale to large graphs, but they may not reproduce the metamorphosis coefficient. Both Newman *et al.* [20] and Guillaume and Latapy [3] propose bipartite configuration models that *sample* node degrees from a distribution (one per partition), create stubs for each node equal to its degree, and then match the stubs. Our bipartite CL model is very similar, and we discuss the connections further in Section 4.2. Guillaume and Latapy [3] considered the clustering coefficients of the product graph, but this may be viewed as being a direct consequence of matching the bipartite degree distribution (a collapsed node of degree d creates a clique of size d in the product graph) and so is different than the metamorphosis condition considered here. The second class are models that are more sophisticated but do not scale to large graphs. On the theoretical side, Kannan *et al.* [21] consider the convergence of a Markov chain rewiring algorithm for generating a bipartite graph and show that their process is rapidly mixing for regular bipartite graphs, that is, graphs where all the node degrees are the same. Later, this result was extended to graphs in which only one partition is required to be regular [22]. In statistics, researchers consider the problem of generating *binary contingency tables* with given row and columns sums. This problem is equivalent to specifying the degrees in a bipartite graph. In [23, 24], the focus is on an algorithm that is *guaranteed* to produce a specified row and column sums, assuming its realizable. The proposed algorithm is not considered practical but rather of theoretical interest. As for one-way graphs, there are also *incremental growth* models in which the output graph is iteratively constructed by adding vertices and edges according to some rule. The methods are sequential since the connections created for new nodes usually depend on the state of the graph at the iteration in which the nodes are created. Guillaume and Latapy [3] analyse a bipartite growing model in which new vertices are linked via a preferential attachment process. Other work has considered a similar idea except that one vertex partition remains fixed while the other grows [25, 26]. Lastly, we note a variety of application-specific generative bipartite graph models have been introduced to model specific networks, including pollination networks in ecology [27, 28], and protein-domain networks in

TABLE 1 *Real-world bipartite graphs*

Name	Partition 1	Partition 2	Edges
CondMat [30, 31]	16,726 authors	22,016 papers	58,595
IMDB [4, 32]	127,823 actors	383,640 movies	1,470,418
Flickr [33, 34]	1,728,701 users	103,648 groups	8,545,307
MovieLens [35, 36]	65,133 movies	71,567 critics	10,000,054
MillionSong [37–39]	1,019,318 users	384,546 songs	48,373,586
Peer2Peer [4, 32]	1,986,588 peers	5,380,546 files	55,829,392
LiveJournal [33, 34]	5,284,451 users	7,489,296 groups	112,307,385

biology [29]. We stress that the biggest difference between bipartite BTER and the models in the second group is scalability.

2. Data sets

We test our methods on publicly available real-world datasets, whose properties are summarized in Table 1. Their degree distributions are shown in Figs 3–5. **CondMat** represents an author–paper network from arXiv preprints in condensed matter physics from 1995–1999 [30]; this has mostly been used in the context of the coauthorship graph, but here we consider the underlying data [31]. The majority of authors have only 1 paper, whereas the most prolific author has 116 papers. Conversely, the most coauthors on a single paper is 18, and the most likely scenario is for a paper to have 2 or 3 authors. **IMDB** links movies and the actors that appeared in them [4, 32], as collected from the Internet Movie Database. The busiest actor was in 294 movies; conversely, the largest production had 646 actors. **Flickr** [33, 34] is an online photo-sharing site, and the network represents group membership of various users. The most connected user is in 2186 groups, whereas the largest group has 34989 members. **MovieLens** [35, 36] is a very famous dataset that links movies and their reviewers/critics. The most-reviewed movie had 34864 reviews, and the most active critic reviewed 7359 movies. This dataset apparently excludes critics with less than 20 reviews. The **MillionSong** [37–39] dataset connects users and the songs they played. The dataset only includes listeners of 10 songs or more. The widest ranging user listened to 4,400 distinct songs. The top song was played by 110,479 distinct listeners. The **Peer2Peer** dataset [4, 32] links users (peers) and the files they uploaded or downloaded. The busiest user touched 19,496 files. On the other hand, the most popular file only had 3396 downloads. **LiveJournal** [33, 34] represents user–group memberships from a blogging site. The most engaged user is in 300 groups, which appears to be the maximum allowed, since there are 5 persons in that category. The largest group has over one million members.

3. Notation

We set up the notation for one-way and two-way graphs in Table 2. We assume all graphs are *simple*, meaning that there are no multiple edges. In the one-way case, we let $n = |V|$ and index nodes in V by $i, j \in \{1, \ldots, n\}$. In the two-way case, we let $n^u = |U|$ and $n^v = |V|$ denote the sizes of partitions one (left) and two (right), respectively. We use $i \in \{1, \ldots, n^u\}$ and $j \in \{1, \ldots, n^v\}$ to index nodes in partition one and two, respectively. Without loss of generality, indexing by i assumes partition one and likewise

TABLE 2 *Notation*

One-way graph	Two-way graph
$G = (V, E)$	$G = (U, V, E)$
Vertices : V	Vtx. Partition 1 : U
	Vtx. Partition 2 : V
Edges : $E \subseteq V \otimes V$	Edges : $E \subseteq U \otimes V$
# Vertices : $n = \|V\|$	# Vertices in U : $n^u = \|U\|$
	# Vertices in V : $n^v = \|V\|$
# Edges : $m = \|E\|$	# Edges : $m = \|E\|$
# Wedges : $n^{\wedge}$	# Caterpillars : n^{Σ}
# Triangles : $n^{\triangle}$	# Butterflies : $n^{⧖}$
Clust. Coeff. : $c = 3n^{\triangle}/n^{\wedge}$	Meta. Coeff. : $c = 4n^{⧖}/n^{\Sigma}$
Vertex Index : $i \in \{1, \ldots, n\}$	Index in U : $i \in \{1, \ldots, n^u\}$
	Index in V : $j \in \{1, \ldots, n^v\}$
Degree of i : $d_i = \|\{j \in V \mid (i,j) \in E\}\|$	Degree of i : $d_i^u = \|\{j \in V \mid (i,j) \in E\}\|$
	Degree of j : $d_j^v = \|\{i \in U \mid (i,j) \in E\}\|$

for j and partition two. For instance, if $i = j = 2$ in the two-way case, although vertices i and j have the same vertex label of 2, they belong to different partitions and are thus two distinct vertices.

4. Fast bipartite CL model

We adapt the CL generative model [17–19] to bipartite graphs and demonstrate its ability to reproduce bipartite degree distributions. We follow the notation described in Section 3.

4.1 *CL for one-way graphs*

Consider a one-way graph $G = (V, E)$. The CL model attempts to match the desired degrees $\{d_1, \ldots, d_n\}$, where d_i denotes the desired degree of vertex i. The model generates a random graph on n vertices such that the probability that vertex i is adjacent to j is given by

$$\Pr((i,j) \in E) = \frac{d_i d_j}{2m}, \quad \text{where} \quad m = \frac{1}{2}\sum_{i=1}^{n} d_i = \text{ desired number of edges.}$$

To ensure the quantities are all true probabilities, we assume $d_i \leq \sqrt{2m}$ for all i. A classical implementation of the CL model on n vertices flips a coin for each of the $\binom{n}{2} = \Omega(n^2)$ possible edges. Many real-world graphs are large and sparse, that is, the number of edges $m = O(n)$. For this reason, we favour 'fast' CL that flips only $2m$ coins [15, 40]. We explain the fast method below in the context of two-way graphs.

4.2 *CL for two-way graphs*

Consider the bipartite graph $G = (U, V, E)$. Here we have separate desired degrees for the vertices in U and V, denoted

$$\{d_i^u\}_{i=1}^{n^u} \quad \text{and} \quad \{d_j^v\}_{j=1}^{n^v},$$

respectively. Necessarily, the sums of the degrees in each partition must be equal to each other and to the number of edges, i.e.,

$$m = \sum_{i=1}^{n^u} d_i^u = \sum_{j=1}^{n^v} d_j^v.$$

Hence, the bipartite CL model generates a random bipartite graph on n^u vertices in the first partition and n^v vertices in the second partition such that

$$\Pr((i,j) \in E) = \frac{d_i^u d_j^v}{m}. \tag{4.1}$$

To ensure these are true probabilities, we assume $d_i^u \leq \sqrt{m}$ for all i and $d_j^v \leq \sqrt{m}$ for all j. A naïve implementation of bipartite CL would flip a coin for all $n^u n^v$ possible edges. Instead, we adopt the 'fast' approach as follows. Rather than flipping a coin for every possible edge, we instead randomly choose two endpoints for every expected edge. Since the graph is sparse, we may assume that $m \ll n^u n^v$, so this approach requires many fewer random samples. For each of the m edges, we choose endpoints in U and V proportional to

$$\Pr(i) = \frac{d_i^u}{m} \quad \text{and} \quad \Pr(j) = \frac{d_j^v}{m},$$

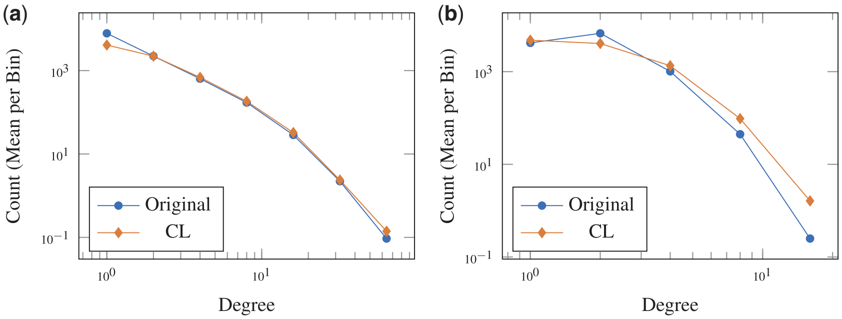
respectively. The probability that edge (i,j) exists is then given by

$$\Pr((i,j) \in E) = m \cdot \Pr(i) \cdot \Pr(j) = \frac{d_i^u d_j^v}{m},$$

which is the same as (4.1). Although both implementations of CL yield identical expected degrees, a key distinction is that multiple edges between the same pair of vertices are possible in fast CL. In practice, for large graphs with heavy-tailed degree distributions, this rarely presents a problem. The fast bipartite CL algorithm is described in Algorithm 1.

Algorithm 1 Fast bipartite CL

```
procedure FBCL($\{d_i^u\}$, $\{d_j^v\}$)
    $m \leftarrow \sum_i d_i^u$
    $E \leftarrow \emptyset$
    for $k = 1, \ldots, m$ do
        Randomly select $i \in U$ proportional to $d_i^u/m$
        Randomly select $j \in V$ proportional to $d_j^v/m$
        $E \leftarrow E \cup (i,j)$                    ▷ Duplicate edges discarded
    end for
return $E$
end procedure
```



Mode 1: CondMat (Authors/Papers)

Mode 2: CondMat (Authors/Papers)

FIG. 3. Degree distributions illustrating the original (circles) versus bipartite fast Chung-Lu (diamonds). The data is log-binned.

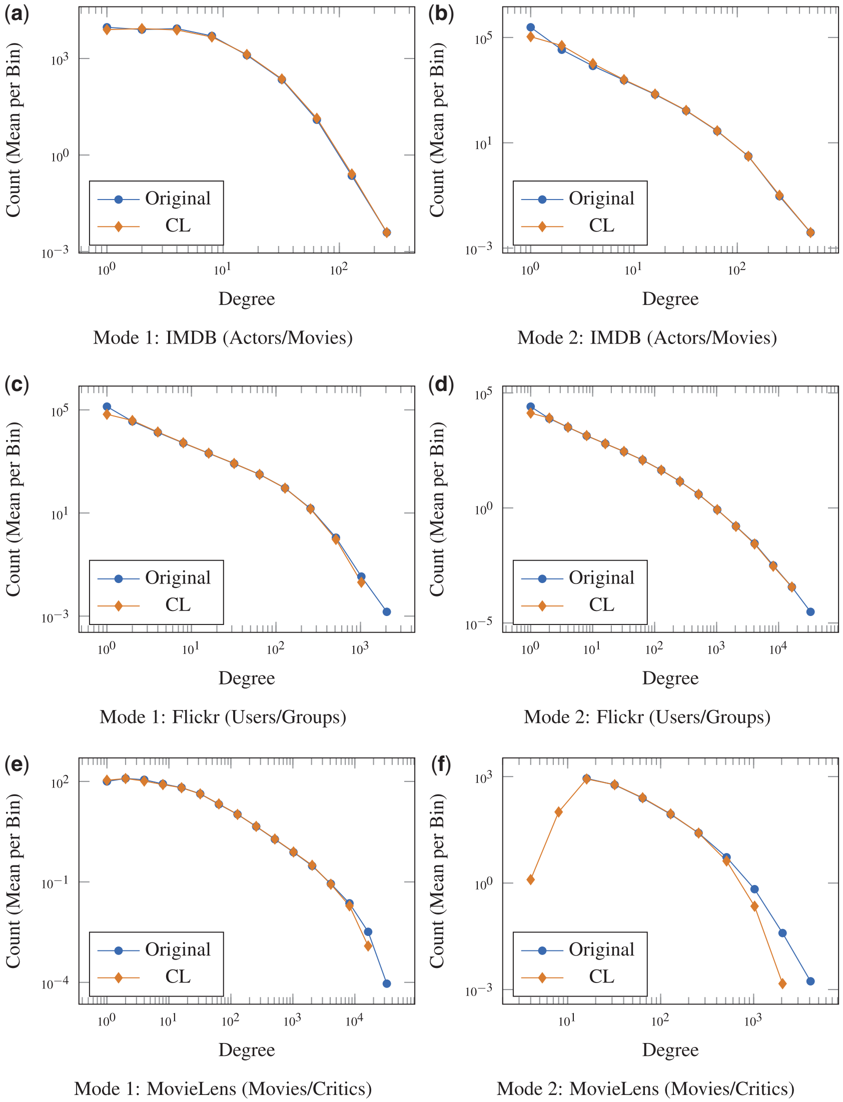
As mentioned in Section 1, a closely related approach is the bipartite configuration model [3, 20]. These differ in some details; for instance, in the CL model, the degree distribution is not specified exactly but rather each degree is sampled from a distribution. However, if we ignore that detail, we may consider that each node $i \in U$ has d_i^u stubs and likewise each node $j \in V$ has d_j^v stubs, and the stubs from partition one are randomly connected to the stubs in partition 2. As with fast bipartite CL, we discard any repeated edges. In all three cases, bipartite CL, fast bipartite CL and the bipartite configuration models, the expected degree of a vertex is the same.

4.3 *Experimental results*

We generate random graphs using CL and the degree distributions of the graphs described in Section 2. The degree distributions are shown in Figs 3–5. The degree distribution of the original graph is shown in blue, and the degree distribution of the graph generated by bipartite CL is shown in orange. These are *binned* degree distributions, as advocated in [41]. We use powers of two for the bin borders, so the x-coordinate 2^k corresponds to the bin from $[2^k, 2^{k+1})$. The y-coordinate is the average value for that bin, including zero values, so the y-coordinate can be less than one. Overall, the degree distributions are very close, especially for IMDB (Fig. 4a and b), Flickr (Fig. 4c and d) and LiveJournal (Fig. 5e and f). For CondMat (Fig. 3a and b), the distribution of degrees on the author nodes is a close match, but there is some trouble matching the paper degree distribution. The model slightly overestimates at the higher end of the degree scale and underestimates at the lower end. This is largely due to the small size of the graph and the very small distribution (maximum of 18 authors for a paper). For MovieLens (Fig. 4e and f), the model generates a few ‘critic’ nodes of degree less than 20, even though no nodes exist in the true degree distribution. A similar phenomenon occurs for the ‘user’ nodes in MillionSong. In general, the CL model cannot handle gaps in the degree distribution because of Poisson distributions in its expectations. In Peer2Peer (Fig. 5c and d), the number of degree-1 peer nodes is underestimated, for reasons described in [40].

4.4 *Shortcomings of bipartite CL*

Overall, if we provide the degree distributions of a real-world graph, the fast bipartite CL model generates a random graph whose degree distribution closely matches the degree distribution of the original graph. However, the graphs generated by CL typically have many fewer butterflies than the original graphs.



Mode 1: IMDB (Actors/Movies)

Mode 2: IMDB (Actors/Movies)

Mode 1: Flickr (Users/Groups)

Mode 2: Flickr (Users/Groups)

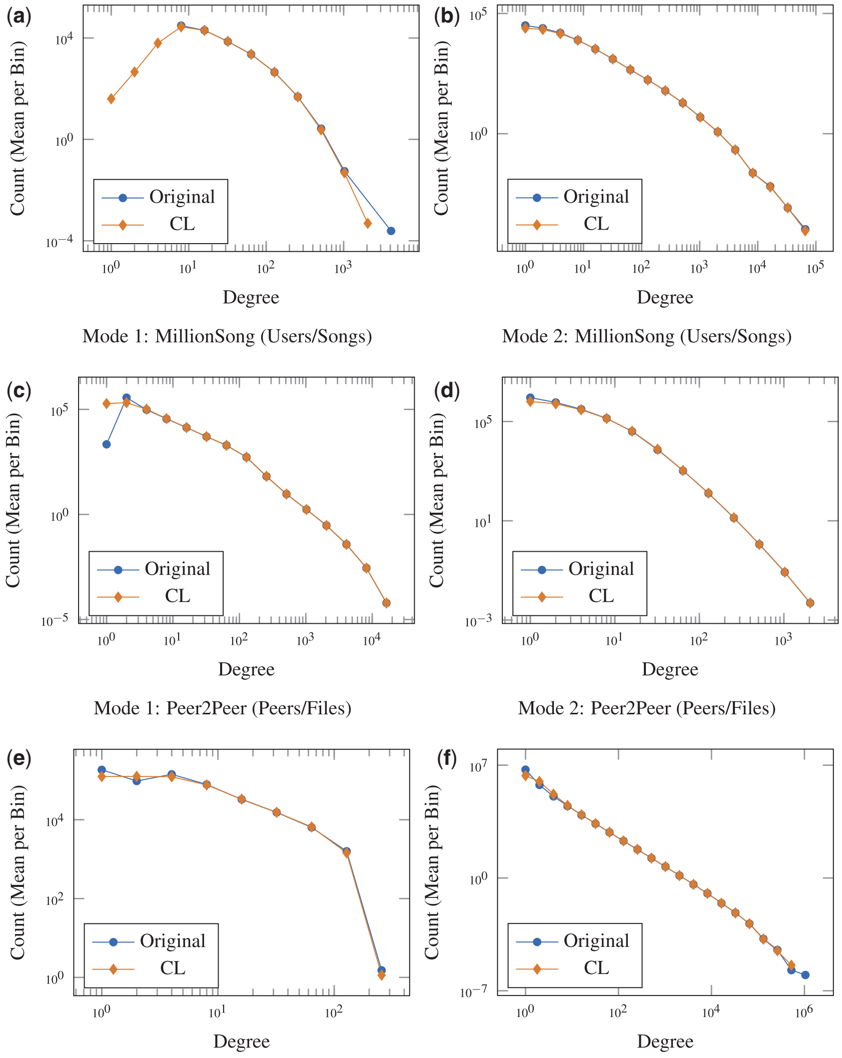
Mode 1: MovieLens (Movies/Critics)

Mode 2: MovieLens (Movies/Critics)

FIG. 4. Degree distributions illustrating the original (circles) versus bipartite fast Chung-Lu (diamonds). The data is log-binned.

Table 3 shows that the number of butterflies from bipartite CL is smaller than the original graph in every case, and sometimes by one or more orders of magnitude (CondMat, IMDB, Peer2Peer). The number of caterpillars for the original and the generated graphs on the other hand, is closer, so the metamorphosis of the original graph is much higher than for the generated graph with low-butterfly counts. This indicates that the bipartite CL model is omitting some important structure.

Note that the butterfly structure is what underlies the cohesive, close-knit structure in many real-world graphs. Just as a triangle can be considered as the smallest unit of cohesion on one-way graphs, butterflies can be considered as the smallest unit of cohesion in bipartite graphs. Conversely, without butterflies,



Mode 1: MillionSong (Users/Songs)

Mode 2: MillionSong (Users/Songs)

Mode 1: Peer2Peer (Peers/Files)

Mode 2: Peer2Peer (Peers/Files)

Mode 1: LiveJournal (Users/Groups)

Mode 2: LiveJournal (Users/Groups)

FIG. 5. Degree distributions illustrating the original (circles) versus bipartite fast Chung-Lu (diamonds). The data is log-binned.

a bipartite graph will not have a community structure. This is our motivation for considering a more complex model in the next section.

5. Bipartite BTER model

In the BTER model for one-way graphs, the goal is to match both the degree distribution as well as the degreewise clustering coefficients [13, 15]. Our goal here is to extend those notions to the bipartite case. We use (1.1) as the bipartite definition of the clustering coefficient, though other metrics exist as surveyed by Latapy *et al.* [4] and Opsahl [8].

TABLE 3 *Properties of the original and the bipartite CL graphs*

Graph	Size		Edges	Cats.	Buts.	Meta.
	n^u	n^v	m	n^{Σ}	$n^{\bowtie}$	c
CondMat-Orig	1.67e4	2.20e4	5.86e4	1.24e6	7.05e4	2.28e−1
CondMat-CL	1.67e4	2.20e4	5.86e4	2.22e6	3.57e2	6.43e−4
IMDB-Orig	1.28e5	3.84e5	1.47e6	8.56e8	3.50e6	1.64e−2
IMDB-CL	1.28e5	3.84e5	1.47e6	1.11e9	1.41e5	5.10e−4
Flickr-Orig	1.73e6	1.04e5	8.55e6	2.57e+12	3.53e+10	5.49e−2
Flickr-CL	1.73e6	1.04e5	8.39e6	2.20e+12	1.52e+10	2.78e−2
MovieLens-Orig	6.51e4	7.16e4	1.00e7	2.46e+13	1.20e+12	1.95e−1
MovieLens-CL	6.51e4	7.16e4	8.78e6	1.37e+13	5.34e+11	1.56e−1
MillionSong-Orig	1.02e6	3.85e5	4.84e7	2.21e+13	2.15e+11	3.89e−2
MillionSong-CL	1.02e6	3.85e5	4.81e7	2.59e+13	6.74e+10	1.04e−2
Peer2Peer-Orig	1.99e6	5.38e6	5.58e7	8.18e+11	3.80e9	1.86e−2
Peer2Peer-CL	1.99e6	5.38e6	5.58e7	1.20e+12	1.14e8	3.79e−4
LiveJournal-Orig	5.28e6	7.49e6	1.12e8	3.36e+14	3.30e+12	3.92e−2
LiveJournal-CL	5.28e6	7.49e6	1.11e8	3.31e+14	2.04e+12	2.47e−2

5.1 *Degreewise metamorphosis coefficient*

BTER matches the degreewise clustering coefficient, so we need a similar measure for bipartite graphs. We describe the degreewise metamorphosis coefficient, which provides a more nuanced measurement of bipartite community structure than the metamorphosis coefficient. To the best of our knowledge, this idea has not been proposed before.

We define the metamorphosis of an edge (i,j) as

$$c_{(i,j)} = \begin{cases} \dfrac{n^{\bowtie}_{(i,j)}}{n^{\Sigma}_{(i,j)}} = \dfrac{\text{number of butterflies containing } (i,j)}{\text{number of caterpillars centered at } (i,j)} & \text{if } n^{\Sigma}_{(i,j)} > 0, \\ 0 & \text{if } n^{\Sigma}_{(i,j)} = 0. \end{cases} \tag{5.1}$$

We know the number of caterpillars centered at (i,j) immediately from the degrees of its endpoints, that is

$$n^{\Sigma}_{(i,j)} = (d^u_i - 1)(d^v_j - 1). \tag{5.2}$$

From this, we define the metamorphosis coefficients of vertices $i \in U$ and $j \in V$ as the mean value over all edges incident to the vertex:

$$c^u_i = \frac{1}{d^u_i} \sum_{(i,j)\in E} c_{(i,j)} \quad \text{and} \quad c^v_j = \frac{1}{d^v_j} \sum_{(i,j)\in E} c_{(i,j)}. \tag{5.3}$$

We may consider (5.3) to be the bipartite analogue of the clustering coefficient of a vertex, as introduced by Watts and Strogatz [42]. Finally, we can define the per-degree metamorphosis coefficients to be

$$c^u_d = \frac{1}{|U_d|}\sum_{i\in U_d} c^u_i \quad \text{and} \quad c^v_d = \frac{1}{|V_d|}\sum_{j\in V_d} c^v_j, \tag{5.4}$$

where U_d and V_d denote the subsets of degree-d nodes, that is

$$U_d = \{i \in U | d^u_i = d\} \quad \text{and} \quad V_d = \{j \in V | d^v_j = d\}.$$

Degreewise metamorphosis coefficients control for the effects of both vertex mode and vertex degree on bipartite clustering. Accordingly, this metric may reveal insights otherwise lost in metrics based only on the ratio of total butterfly to caterpillar counts. To illustrate, consider the CondMat author–paper network, whose degreewise metamorphosis coefficients are shown by the blue line in Fig. 6c and d. The degreewise metamorphosis coefficients are also binned in the same way that we binned the data for the degree distributions: We use powers of two for the bins, so the x-coordinate 2^k corresponds to the bin from $[2^k, 2^{k+1})$. The y-coordinate is the average value for that bin. For the binning, we define $c^u_d = 0$ for any degree such that $|U_d| = 0$ (i.e., when there are no nodes of degree d), and likewise for c^v_c. We see that degreewise metamorphosis coefficients in the author mode (Fig. 6c) are higher for low degrees than for high degrees, meaning that authors with fewer papers tended to have higher proportion of repeat collaborations than authors with many papers. Conversely, the paper mode (Fig. 6d) shows that authors of papers with few authors tend to have more repeats of the same author set.

5.2 *Affinity blocks*

In BTER for one-way graphs, dense ER subgraphs are key to producing triangles. For bipartite BTER, we will use dense bipartite ER subgraphs to produce butterflies. We refer to these dense ER subgraphs as *affinity blocks*.

In (one-way) BTER, an affinity block ideally consists of a set of $d + 1$ vertices of degree d. The connectivity of each block is computed according to the degree-d clustering coefficient. The bipartite affinity blocks are similar in spirit, but the bipartite nature of the graph raises issues that require an entirely new approach to the block construction.

The first key difference is that each affinity block in bipartite BTER consists of *two* sets of vertices, one from each partition. While each partition set in an affinity block ideally contains vertices of the same degree, the degrees do not necessarily match between partition sets. Indeed, in many bipartite graphs, one partition set may have a very different range of degrees than the other, so attempting to create blocks that match inter-partition degree is not a realistic goal.

Consequently, a second key difference concerns how we determine the *sizes* of the partition sets for the affinity blocks. In the one-way BTER method, the size of each block only depends on the degree of the vertices in the block. In the two-way case, the sizes of each partition set within a block are more complicated.

To work out the calculations of sizes and connectivity for the blocks, we consider building a single affinity block denoted by $\hat{G} = (\hat{U}, \hat{V}, \hat{E})$. Without loss of generality, we assume all nodes in $\hat{U}$ want degree $\hat{d}^u$ and all nodes in $\hat{V}$ want degree $\hat{d}^v$. Note that these degrees are with respect to the entire graph, not the subgraph. We further assume all nodes in $\hat{U}$ want metamorphosis coefficient $\hat{c}^u$ and likewise for $\hat{V}$ and $\hat{c}^v$. When matching a real-world graph, we choose the degreewise metamorphosis coefficients corresponding to the target degrees as defined in (5.4), that is,

$$\hat{c}^u = c_{\hat{d}^u} \quad \text{and} \quad \hat{c}^v = c_{\hat{d}^v}.$$

The goal is to determine the sizes $\hat{n}^u = |\hat{U}|$ and $\hat{n}^v = |\hat{V}|$ and the connectivity, ρ, which is the probability of an edge between a node in $\hat{n}^u$ and a vertex in $\hat{n}^v$, and thus the number of edges, $|\hat{E}|$

For $i \in \hat{U}$, we can compute its expected metamorphosis coefficient as follows. By definition (5.3), we have

$$c_i^u = \frac{1}{d_i^u} \sum_{(i,j)\in E} c_{(i,j)} = \frac{1}{d_i^u} \left(\sum_{(i,j)\in\hat{E}} c_{(i,j)} + \sum_{(i,j)\in E\setminus\hat{E}} c_{(i,j)} \right) \approx \frac{1}{d_i^u} \sum_{(i,j)\in\hat{E}} c_{(i,j)}. \tag{5.5}$$

The last step comes from the assumption that nearly all butterflies in the larger graph come from the affinity blocks. Using definition (5.1), we can rewrite (5.5) as

$$c_i^u = \frac{1}{d_i^u} \sum_{(i,j)\in\hat{E}} \frac{n_{(i,j)}^{\text{⧖}}}{n_{(i,j)}^{\Sigma}} = \frac{1}{d_i^u} \sum_{(i,j)\in\hat{E}} \frac{n_{(i,j)}^{\text{⧖}}}{(d_i^u - 1)(d_j^v - 1)}. \tag{5.6}$$

We have assumed that the degrees and clustering coefficients within $\hat{G}$ are constant, so (5.6) becomes

$$\hat{c}^u = \frac{1}{\hat{d}^u(\hat{d}^u - 1)(\hat{d}^v - 1)} \sum_{(i,j)\in\hat{E}} n_{(i,j)}^{\text{⧖}} = \frac{2n_i^{\text{⧖}}}{\hat{d}^u(\hat{d}^u - 1)(\hat{d}^v - 1)}. \tag{5.7}$$

The last equality uses the fact that the sum of butterflies involving edges of the form (i,j) is equal to two times the number of butterflies involving node i since each such butterfly has two edges involving i. In expectation,

$$n_i^{\text{⧖}} \doteq \rho^4(\hat{n}^u - 1)\binom{\hat{n}^v}{2}, \tag{5.8}$$

because there are $(\hat{n}^u - 1)$ other choices for the second node in $\hat{U}$ and $\binom{\hat{n}^v}{2}$ choices for the two nodes in $\hat{V}$. Finally, ρ^4 is the probability that all four edges exist. Combining (5.7) and (5.8) gives

$$\hat{c}^u \doteq \frac{\rho^4\hat{n}^v(\hat{n}^u - 1)(\hat{n}^v - 1)}{\hat{d}^u(\hat{d}^u - 1)(\hat{d}^v - 1)}. \tag{5.9}$$

Using analogous reasoning for $\hat{c}^v$, we ultimately want

$$\rho^4 = \frac{\hat{c}^u\hat{d}^u(\hat{d}^u - 1)(\hat{d}^v - 1)}{\hat{n}^v(\hat{n}^u - 1)(\hat{n}^v - 1)} = \frac{\hat{c}^v\hat{d}^v(\hat{d}^u - 1)(\hat{d}^v - 1)}{\hat{n}^u(\hat{n}^u - 1)(\hat{n}^v - 1)}. \tag{5.10}$$

Ideally, we would have

$$\hat{n}^u = \hat{d}^v \quad \text{and} \quad \hat{n}^v = \hat{d}^u, \tag{5.11}$$

but we cannot generally satisfy (5.10) and (5.11) at the same time. Therefore, we choose one of the equalities in (5.11) and then solve for the other number of nodes and connectivity using (5.10) to get

$$\hat{n}^u = \hat{d}^v \quad \Rightarrow \quad \hat{n}^v = \frac{\hat{c}^u}{\hat{c}^v}\hat{d}^u, \quad \rho = \frac{(\hat{d}^u - 1)(\hat{c}^v)^2}{\hat{c}^u\hat{d}^u - \hat{c}^v}, \tag{5.12}$$

$$\hat{n}^v = \hat{d}^u \quad \Rightarrow \quad \hat{n}^u = \frac{\hat{c}^v}{\hat{c}^u}\hat{d}^v, \quad \rho = \frac{(\hat{d}^v - 1)(\hat{c}^u)^2}{\hat{c}^v\hat{d}^v - \hat{c}^u}. \tag{5.13}$$

So that we can have the possibility of using a complete bipartite subgraph to yield metamorphosis coefficients of one, we constrain our choices to satisfy:

$$\hat{n}^u \geq \hat{d}^v \quad \text{and} \quad \hat{n}^v \geq \hat{d}^u. \tag{5.14}$$

To satisfy (5.14), we choose (5.12) if $\frac{\hat{c}^u}{\hat{c}^v} \geq 1$ and (5.13) otherwise. It is easy to see from (5.10) that this has the added bonus of ensuring $\rho \leq 1$. This logic forms the basis of building affinity blocks for bipartite BTER.

5.3 *Bipartite BTER algorithm*

As explained in Section 5.2, the affinity block building process for bipartite BTER takes into account both the desired vertex degrees and corresponding desired per-degree metamorphosis coefficients when setting the partition sizes and connectivity of each affinity block. This information is then used to model each affinity block as an ER subgraph, where the preponderance of the graph butterflies are created. Accordingly, the affinity block construction is key to matching per-degree metamorphosis coefficients. In order to match the desired degree distribution, the remaining *excess degree* (i.e., edges not used in constructing the affinity blocks) is connected via a fast bipartite CL procedure. The full bipartite BTER algorithm is listed in Algorithm 2.

5.4 *Experimental results*

We generate bipartite BTER graphs using the procedure described in Section 5.3. We first discuss a single graph: CondMat. We show the resulting degree distribution and degreewise metamorphosis coefficients in Fig. 6. The degree distributions, shown in Fig. 6a and b, show little difference between bipartite BTER and CL; both are good matches to the original degree distribution. The degreewise metamorphosis coefficients are shown in Fig. 6c and d. Although there is not a perfect match between bipartite BTER and the original graph, it is much better than CL, which has almost no butterflies.

Summary data for all graphs is shown in Table 4. This is the same as Table 3 except that now we have added a row for bipartite BTER. For the first four graphs, we are reporting average values over 100 experiments, and we also report the entire range of values; we do not do multiple trials for the larger graphs because the postprocessing to count the butterflies is extremely expensive. The numbers of butterflies and metamorphosis coefficients are significantly improved as compared to CL. We see that CondMat has 70,000 butterflies, bipartite CL produces less than 400 butterflies, but bipartite BTER produces 120,000 butterflies. The bipartite BTER number is a slight overestimate, but overall much better than CL.

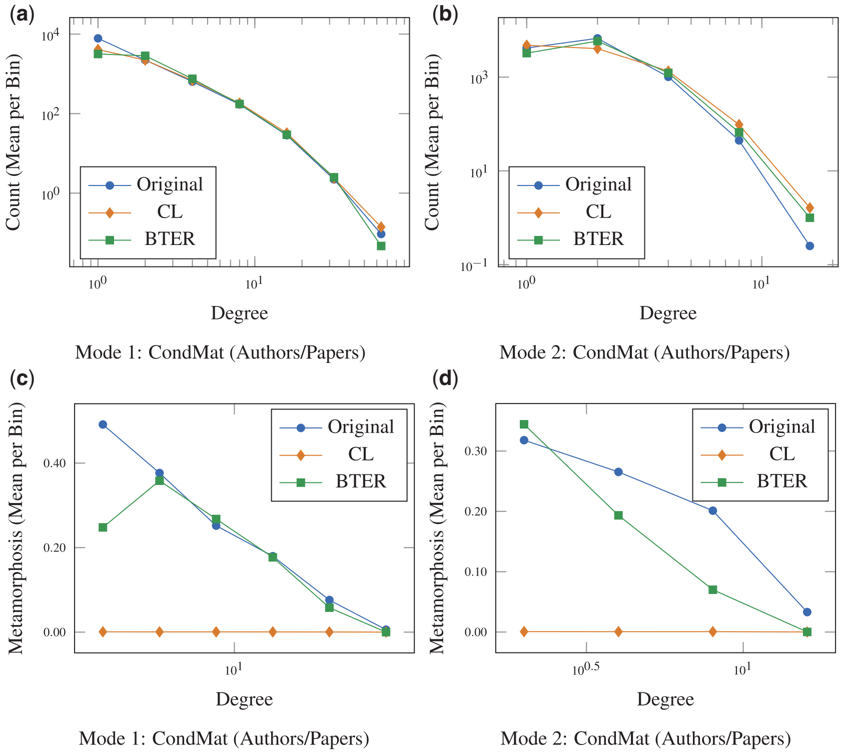
For the first two graphs, where we see a large difference between the metamorphosis coefficients, we consider the impact on the product graphs ($A'A$ and AA' where A is the adjacency matrix) in Fig. 7. We

Algorithm 2 Bipartite BTER

```
procedure BIBTER($\{d^u_i\}$, $\{d^v_j\}$, $\{c^u_d\}$, $\{c^v_d\}$ )
We assume that degrees are sorted in increasing order
    $m \leftarrow \sum_i d^u_i$
    $E \leftarrow \emptyset$
    $\{e^u_i\} \leftarrow \{d^u_i\}, \{e^v_j\} \leftarrow \{d^v_j\}$                    ▷ Excess degree initialization
    $i \leftarrow \min\{i | d^u_i > 1\}, j \leftarrow \min\{j | d^v_j > 1\}$
    repeat                                            ▷ Create affinity blocks until nodes exhausted
        $\hat{d}^u \leftarrow d^u_i, \hat{d}^v \leftarrow d^v_j$
        $\hat{c}^u \leftarrow c_{\hat{d}^u}, \hat{c}^v \leftarrow c_{\hat{d}^v}$
        if $\hat{c}^u / \hat{c}^v \geq 1$ then
            $\hat{n}^u \leftarrow \hat{d}^v, \hat{n}^v \leftarrow \mathtt{round}\left(\frac{\hat{c}^u}{\hat{c}^v}\hat{d}^u\right), \rho \leftarrow \left(\frac{(\hat{d}^u - 1)(\hat{c}^v)^2}{\hat{c}^u \hat{d}^u - \hat{c}^v}\right)^{1/4}$
        else
            $\hat{n}^v \leftarrow \hat{d}^u, \hat{n}^u \leftarrow \mathtt{round}\left(\frac{\hat{c}^v}{\hat{c}^u}\hat{d}^v\right), \rho \leftarrow \left(\frac{(\hat{d}^v - 1)(\hat{c}^u)^2}{\hat{c}^v \hat{d}^v - \hat{c}^u}\right)^{1/4}$
        end if
        if $i + \hat{n}^u - 1 \leq n^u$ and $j + \hat{n}^v - 1 \leq n^v$ then                    ▷ Create ER subgraph
            for $\hat{i} = i, i + 1, \ldots, i + \hat{n}^u - 1$ do
                for $\hat{j} = j, j + 1, \ldots, j + \hat{n}^v - 1$ do
                    $r \leftarrow U(0, 1)$                              ▷ Uniform random value in [0, 1]
                    if $r <= \rho$ then
                        $E \leftarrow E \cup (\hat{i}, \hat{j})$
                        $e^u_{\hat{i}} = \max\{0, e^u_{\hat{i}} - 1\}, e^v_{\hat{j}} = \max\{0, e^v_{\hat{j}} - 1\}$          ▷ Update excess degree
                    end if
                end for
            end for
        end if
        $i \leftarrow i + \hat{n}^u, j \leftarrow j + \hat{n}^v$
    until $i > n^u$ or $j > n^v$
    $E \leftarrow E \cup$ FBCL($\{e^u_i\}, \{e^v_j\}$)
end procedure
```

expect that CL will produce too many nonzeros in the product graph because there will not be enough overlap in the cliques. Indeed, CL predicts too many nonzeros, but BTER is much closer to the original.

The degreewise metamorphosis coefficients for the remaining graphs are shown in Figs 8 and 9. Even in cases where bipartite CL's overall numbers are close to the original graph as shown in Table 4, the degreewise metamorphosis coefficients are essentially zero. Bipartite BTER corrects this problem and gets metamorphosis coefficients that are much closer to what we see in the original graph. In some cases like MovieLens (Fig. 8e), the CL metamorphosis coefficients are much higher than zero, which is the result of the overall high density of the graph. Indeed, it has been proven (see Lemma 1 in [43]) that under mild assumptions, *any* bipartite graph with m edges and partition sizes n^u and n^v contains at least on the order of $(\frac{n^u}{n^v})^2 \cdot (\frac{m}{n^u+n^v})^4$ many butterflies. Thus, the presence of a certain minimum number butterflies in bipartite graphs of sufficient density is inevitable; nevertheless, the original graph still has higher values that are matched better by bipartite BTER. MovieLens also has an unusual profile in Mode 2



Mode 1: CondMat (Authors/Papers)

Mode 2: CondMat (Authors/Papers)

Mode 1: CondMat (Authors/Papers)

Mode 2: CondMat (Authors/Papers)

FIG. 6. Degree distribution and degreewise metamorphosis coefficients on the original CondMat graph as well as the models generated by CL and bipartite BTER.

(see Fig. 8f), which is really just an artefact of the data collection. Nevertheless, bipartite BTER is able to obtain a reasonable approximation. For completeness, the degree distributions for bipartite CL and BTER as compared to the original graphs are shown in Figs 10 and 11. There is little difference between bipartite BTER and CL in terms of the degree distribution.

6. Conclusions

We have considered the problem of how to generate realistic bipartite graphs to reproduce the characteristics of large real-world networks. Our first model, bipartite CL, accurately reproduces the degree distribution. Our second model, bipartite BTER, goes further to capture the degreewise metamorphosis coefficients. High coefficients are indicative of a relatively large number of butterflies indicating cohesion or community structure, which is rare in sparse random graphs unless there are some behaviors that go beyond just the degree structure. Creating realistic graph models leads to some hypotheses about the ways the graphs were formed. In the cases where CL greatly underestimated the number of butterflies (CondMat, IMDB and Peer2Peer), we can surmise that there is some significant community-like behavior. This is easy to see for authorship of papers (CondMat) and actors appearing in movies (IMBD). For the Peer2Peer network, we might hypothesize that some tight-knit peer groups are sharing many files between themselves. The other graphs have a smaller difference between the number of butterflies

TABLE 4 *Metamorphosis coefficients for original, bipartite CL, and bipartite BTER graphs*

Graph	Size		Edges	Cats.	Buts.	Meta.
	n^u	n^v	m	n^{Σ}	$n^{⧖}$	c
CondMat-Orig	1.67e4	2.20e4	5.86e4	1.24e6	7.05e4	2.28e−1
CondMat-CL	1.67e4	2.20e4	5.86e4	2.22e6	3.57e2	6.43e−4
CondMat-BTER	1.67e4	2.20e4	6.00e4	2.36e6	1.10e5	1.86e−1
BTER 100 trial range	[1.67–1.67]	[2.20–2.20]	[6.00–6.01]	[2.31–2.42]	[1.01–1.16]	[1.73–1.99]
IMDB-Orig	1.28e5	3.84e5	1.47e6	8.56e8	3.50e6	1.64e−2
IMDB-CL	1.28e5	3.84e5	1.47e6	1.11e9	1.41e5	5.10e−4
IMDB-BTER	1.28e5	3.84e5	1.47e6	1.35e9	6.77e6	2.00e−2
BTER 100 trial range	[1.28–1.28]	[3.84–3.84]	[1.47–1.47]	[1.35–1.36]	[6.62–6.91]	[1.95–2.04]
Flickr-Orig	1.73e6	1.04e5	8.55e6	2.57e12	3.53e10	5.49e−2
Flickr-CL	1.73e6	1.04e5	8.39e6	2.20e12	1.52e10	2.78e−2
Flickr-BTER	3.96e5	1.04e5	8.34e6	2.36e12	4.25e10	7.21e−2
BTER 100 trial range	[3.96–3.96]	[1.04–1.04]	[8.34–8.34]	[2.35–2.37]	[4.23–4.27]	[7.19–7.23]
MovieLens-Orig	6.51e4	7.16e4	1.00e7	2.46e13	1.20e12	1.95e−1
MovieLens-CL	6.51e4	7.16e4	8.78e6	1.37e13	5.34e11	1.56e−1
MovieLens-BTER	1.07e4	6.99e4	8.52e6	1.23e13	1.08e12	3.51e−1
BTER 100 trial range	[1.07–1.07]	[6.99–6.99]	[8.51–8.52]	[1.23–1.23]	[1.07–1.08]	[3.50–3.51]
MillionSong-Orig	1.02e6	3.85e5	4.84e7	2.21e13	2.15e11	3.89e−2
MillionSong-CL	1.02e6	3.85e5	4.81e7	2.59e13	6.74e10	1.04e−2
MillionSong-BTER	1.02e6	3.85e5	4.80e7	2.93e13	1.90e11	2.59e−2
Peer2Peer-Orig	1.99e6	5.38e6	5.58e7	8.18e11	3.80e9	1.86e−2
Peer2Peer-CL	1.99e6	5.38e6	5.58e7	1.20e12	1.14e8	3.79e−4
Peer2Peer-BTER	1.99e6	5.38e6	5.60e7	1.66e12	6.71e9	1.61e−2
LiveJournal-Orig	5.28e6	7.49e6	1.12e8	3.36e14	3.30e12	3.92e−2
LiveJournal-CL	5.28e6	7.49e6	1.11e8	3.31e14	2.04e12	2.47e−2
LiveJournal-BTER	3.20e6	7.49e6	1.10e8	3.50e14	4.61e12	5.26e−2

produced by CL and the real-world graph, indicating that there is much less community structure. In fact, some graphs are so dense that CL has a nonzero metamorphosis coefficient, especially MovieLens.

Although these models match real-world observations in some aspects, more study is needed to understand their limitations. How well do these models reproduce other metrics such as graph diameter, singular values of the adjacency matrix, centrality measures, joint degree distributions, assortativity, subgraph census, and so on? Can we find evidence that affinity blocks exist in real-world graphs? We have created non-overlapping blocks; is that realistic? We have not mentioned time evolution, but certainly all networks are evolving in time and we need models that capture such changes. We stress that bipartite BTER is not geared towards creating a hierarchy of communities for bipartite graphs as described in [9], but it could potentially be modified to do so by grouping the affinity blocks. These are but a few topics for future investigation.

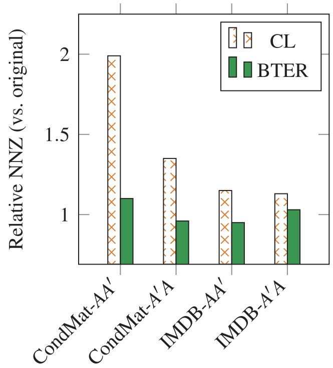
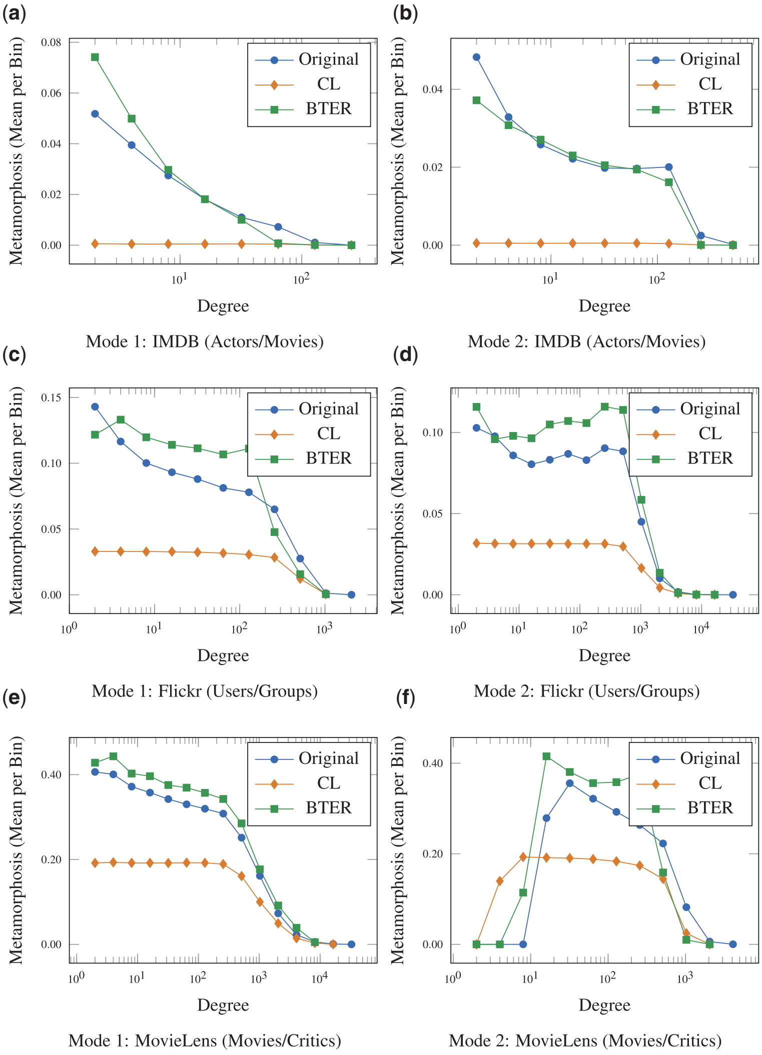


FIG. 7. Comparison of number of nonzeros in product graphs AA' and $A'A$ from CL and BTER as compared to from the original.



Mode 1: IMDB (Actors/Movies)

Mode 2: IMDB (Actors/Movies)

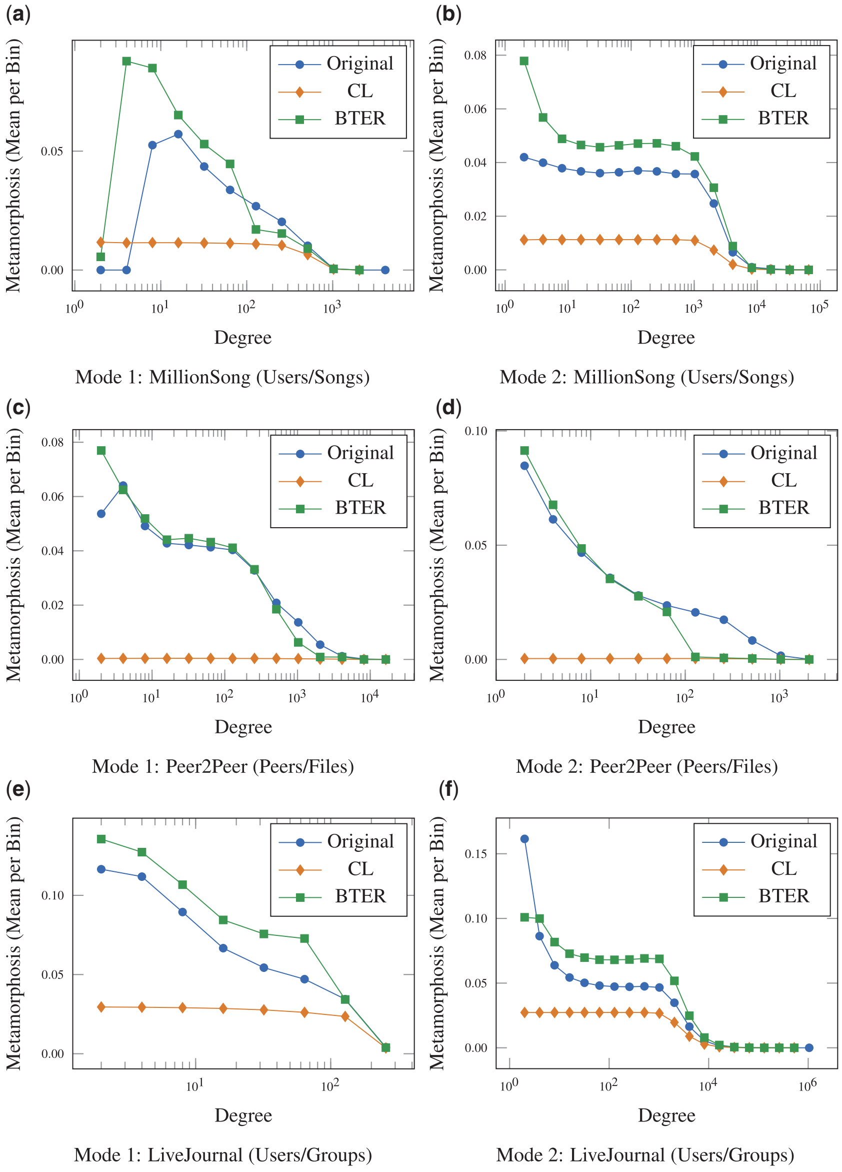
Mode 1: Flickr (Users/Groups)

Mode 2: Flickr (Users/Groups)

Mode 1: MovieLens (Movies/Critics)

Mode 2: MovieLens (Movies/Critics)

FIG. 8. Degreewise metamorphosis coefficient of the original graph (circles), fast bipartite CL (diamonds), and bipartite BTER (squares).



Mode 1: MillionSong (Users/Songs)

Mode 2: MillionSong (Users/Songs)

Mode 1: Peer2Peer (Peers/Files)

Mode 2: Peer2Peer (Peers/Files)

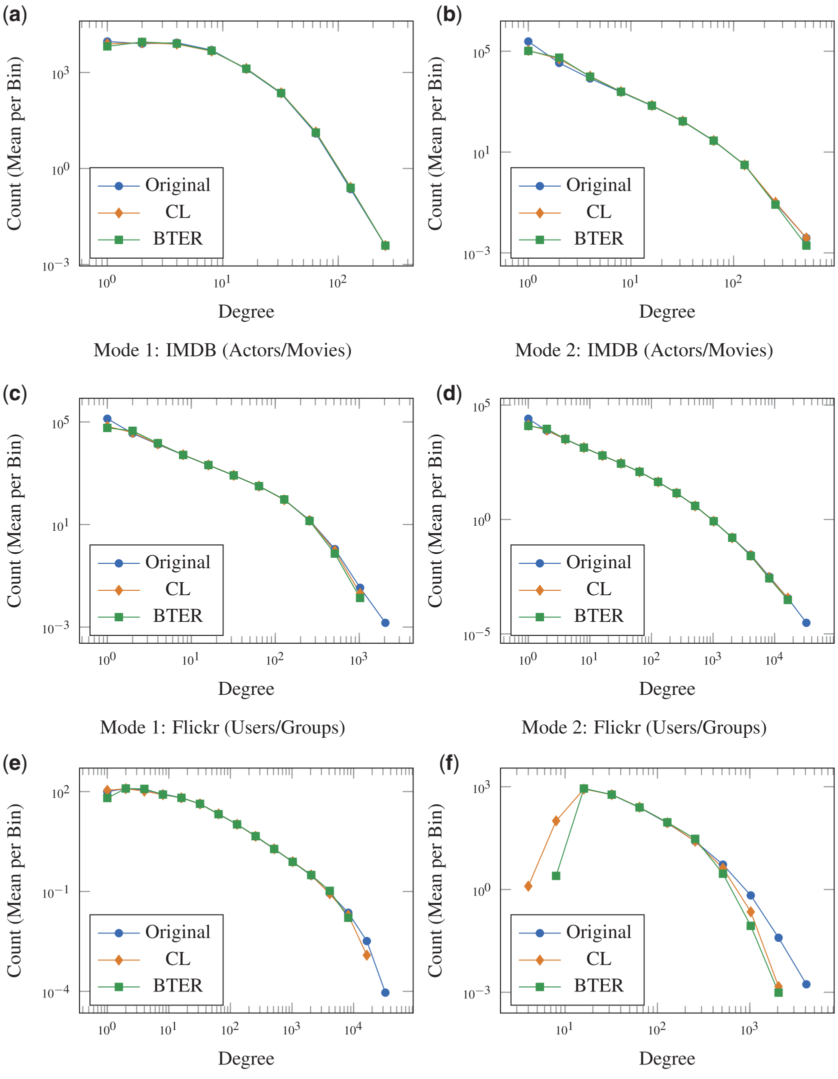
Mode 1: LiveJournal (Users/Groups)

Mode 2: LiveJournal (Users/Groups)

FIG. 9. Degreewise metamorphosis coefficient of the original graph (circles), fast bipartite CL (diamonds), and bipartite BTER (squares).

Acknowledgements

We thank our colleagues for helpful discussions in the course of this work, especially Grey Ballard, Kenny Chowdhary, Danny Dunlavy, Kevin Matulef and Michael Wolf. We also thank the two anonymous referees for insightful comments that improved the work considerably.



Mode 1: IMDB (Actors/Movies)

Mode 2: IMDB (Actors/Movies)

Mode 1: Flickr (Users/Groups)

Mode 2: Flickr (Users/Groups)

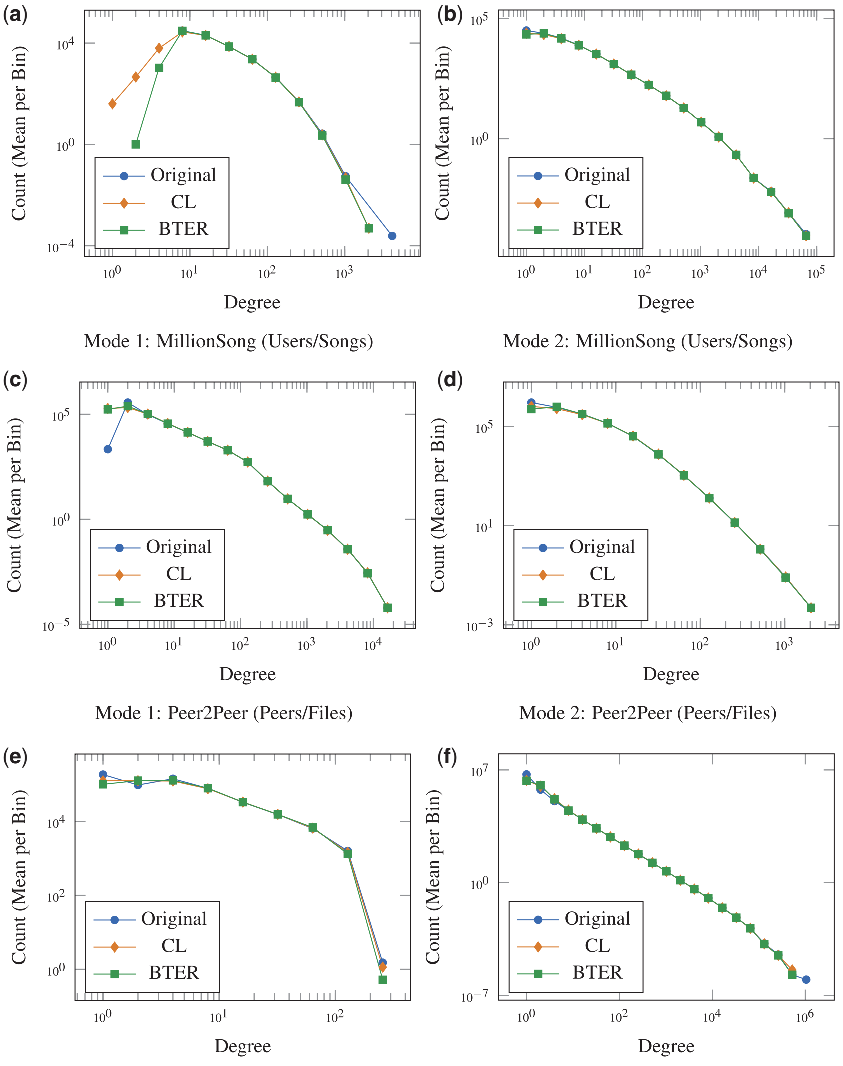
Mode 1: MovieLens (Movies/Critics)

Mode 2: MovieLens (Movies/Critics)

FIG. 10. Degree distributions illustrating the original (circles), fast bipartite CL (diamonds), and bipartite BTER (squares). The data is log-binned.

Funding

This work was supported in part by the Defense Advanced Research Project Agency (DARPA). Sandia National Laboratories is a multi-mission laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.



Mode 1: MillionSong (Users/Songs)

Mode 2: MillionSong (Users/Songs)

Mode 1: Peer2Peer (Peers/Files)

Mode 2: Peer2Peer (Peers/Files)

Mode 1: LiveJournal (Users/Groups)

Mode 2: LiveJournal (Users/Groups)

FIG. 11. Degree distributions illustrating the original (circles), fast bipartite CL (diamonds), and bipartite BTER (squares). The data is log-binned.

REFERENCES

1. DHILLON, I. S. (2001) Co-clustering documents and words using bipartite spectral graph partitioning. In *KDD'01: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, 269–274. New York, NY: ACM.
2. GUILLAUME, J.-L. & LATAPY, M. (2004) Bipartite structure of all complex networks. *Inf. Process. Lett.*, **90**, 215–221.
3. GUILLAUME, J.-L. & LATAPY, M. (2006) Bipartite graphs as models of complex networks. *Physica A*, **371**, 795–813.

4. LATAPY, M., MAGNIEN, C. & VECCHIO, N. D. (2008) Basic notions for the analysis of large two-mode networks. *Social Networks*, **30**, 31–48.

5. ROBINS, G. & ALEXANDER, M. (2004) Small worlds among interlocking directors: network structure and distance in bipartite graphs. *Comput. Math. Organiz. Theory*, **10**, 69–94.

6. SUN, J., QU, H., CHAKRABARTI, D. & FALOUTSOS, C. (2005) Relevance search and anomaly detection in bipartite graphs. *SIGKDD Explor. Newsl.*, **7**, 48–55.

7. HOLLAND, P. W. & LEINHARDT, S. (1970) A Method for detecting structure in sociometric data. *Am. J. Social.*, **76**, 492–513.

8. OPSAHL, T. (2013) Triadic closure in two-mode networks: redefining the global and local clustering coefficients. *Soc. Networks*, **35**, 159–167.

9. SARIYUCE, A. E. & PINAR, A. (2016) Butterfly Effect: Peeling Bipartite Networks. *arXiv:1611.02756*.

10. BARBER, M. J. (2007) Modularity and community detection in bipartite networks. *Phy. Rev. E*, **76**, 066102.

11. MIYAUCHI, A. & SUKEGAWA, N. (2014) Maximizing Barber's bipartite modularity is also hard. *Optim. Lett.*, **9**, 897–913.

12. GUIMERÀ, R., SALES-PARDO, M. & AMARAL, L. A. N. (2007) Module identification in bipartite and directed networks. *Phys. Rev. E*, **76**.

13. SESHADHRI, C., KOLDA, T. G. & PINAR, A. (2012) Community Structure and Scale-free Collections of Erdős-Rényi Graphs. *Phys. Rev. E*, **85**.

14. BARABÁSI, A.-L. & ALBERT, R. (1999) Emergence of scaling in random networks. *Science*, **286**, 509–512.

15. KOLDA, T. G., PINAR, A., PLANTENGA, T. & SESHADHRI, C. (2014) A scalable generative graph model with community structure. *SIAM J. Sci. Comput.*, **36**, C424–C452.

16. NEWMAN, M. E. J. (2003) The structure and function of complex networks. *SIAM Rev.*, **45**, 167–256.

17. AIELLO, W., CHUNG, F. & LU, L. (2001) A random graph model for power law graphs. *Exp. Math.*, **10**, 53–66.

18. CHUNG, F. & LU, L. (2002a) The average distances in random graphs with given expected degrees. *Proc. Nat. Acad. Sci.*, **99**, 15879–15882.

19. CHUNG, F. & LU, L. (2002b) Connected components in random graphs with given degree sequences. *Ann. Comb.*, **6**, 125–145.

20. NEWMAN, M. E. J., STROGATZ, S. H. & WATTS, D. J. (2001) Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, **64**.

21. KANNAN, R., TETALI, P. & VEMPALA, S. (1999) Simple Markov-chain algorithms for generating bipartite graphs and tournaments. *Random Struc. Algorithms*, **14**, 293–308.

22. MIKLÓS, I., ERDŐS, P. L. & SOUKUP, L. (2013) Towards random uniform sampling of bipartite graphs with given degree sequence. *Electronic J. Combinatorics*, **20**, P16.

23. BEZÁKOVÁ, I., BHATNAGAR, N. & VIGODA, E. (2006) Sampling binary contingency tables with a greedy start. *Random Struct. Algorithms*, **30**, 168–205.

24. CHEN, Y., DIACONIS, P., HOLMES, S. P. & LIU, J. S. (2005) Sequential Monte Carlo Methods for Statistical Analysis of Tables. *J. Am. Statistical Assoc.*, **100**, 109–120.

25. Ganguly, N., Ghosh, S., Krueger, T. & Srivastava, A. (2012) Degree distributions of evolving alphabetic bipartite networks and their projections. *Theor. Comput. Sci.*, **466**, 20–36.

26. PERUANI, F., CHOUDHURY, M., MUKHERJEE, A. & GANGULY, N. (2007) Emergence of a non-scaling degree distribution in bipartite networks: a numerical and analytical study. *Europhysics Letters (EPL)*, **79**, 28001.

27. DORMANN, C. F., FRÜND, J., BLÜTHGEN, N. & GRUBER, B. (2009) Indices, graphs and null models: analyzing bipartite ecological networks. *Open Ecol. J.*, **2**, 7–24.

28. SAAVEDRA, S., REED-TSOCHAS, F. & UZZI, B. (2008) A simple model of bipartite cooperation for ecological and organizational networks. *Nature*, **457**, 463–466.

29. NACHER, J. C., OCHIAI, T., HAYASHIDA, M. & AKUTSU, T. (2009) A mathematical model for generating bipartite graphs and its application to protein networks. *J. Phys. A: Math. Theor.*, **42**, 485005.

30. NEWMAN, M. E. J. (2001) The structure of scientific collaboration networks. *Proc. Nat. Acad. Sci.*, **98**, 404–409.

31. OPSAHL, T. (2015) Tnet Datasets.

32. LATAPY, M. (2006) Bipartite Graph Tools, Version 1.0.

33. Mislove, A. (2015) Personal Communication.

34. Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P. & Bhattacharjee, B. (2007) Measurement and analysis of online social networks. In *IMC'07: Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, 29–42, New York, NY: ACM.

35. Davies, R. (2009) MovieLens 10M. GroupLens, Department of Computer Science and Engineering, University of Minnesota. https://grouplens.org/datasets/movielens/10m/ (last accessed February 21, 2017).

36. Harper, F. M. & Konstan, J. A. (2015) The MovieLens datasets. *ACM Trans. Interactive Intell. Syst.*, **5**, 1–19.

37. Bertin-Mahieux, T., Ellis, D. P. W., Whitman, B. & Lamere, P. (2011) The Million Song Dataset. In *ISMIR 2011: Proc. 12th Intl. Conf. on Music Information Retrieval*. (A. Klapuri & C. Leider eds). Miami, FL: University of Miami.

38. McFee, B., Bertin-Mahieux, T., Ellis, D. P. & Lanckriet, G. R. (2012) The million song dataset challenge. In *WWW'12 Companion: Proc. 21st Intl. Conf. Companion on World Wide Web*, New York, NY: ACM. 909–916.

39. The Echo Nest Taste Profile Subset, the Official User Data Collection for the Million Song Dataset. `http://labrosa.ee.columbia.edu/millionsong/tasteprofile` (last accessed May 15, 2015).

40. Durak, N., Kolda, T. G., Pinar, A. & Seshadhri, C. (2013) A Scalable Null model for directed graphs matching all degree distributions: in, out, and reciprocal. In *NSW 2013: Proceedings of IEEE 2013 2nd International Network Science Workshop*, 23–30, IEEE Computer Society 2013, ISBN 978-1-4799-0436-5.

41. Milojević, S. (2010) Power Law Distributions in Information Science: Making the Case for Logarithmic Binning. *J. Am. Soc. Inf. Sci. Technol.*, **61**, 2417–2425.

42. Watts, D. & Strogatz, S. (1998) Collective dynamics of 'small-world' networks. *Nature*, **393**, 440–442.

43. Tait, M. & Verstraëte, J. (2016) On sets of integers with restrictions on their products. *Eur. J. Comb.*, **51**, 268–274.