SAND2016-7228R

Measuring and Modeling Bipartite Graphs with Community Structure

SINAN AKSOY
Department of Mathematics, University of California, San Diego, CA

TAMARA G. KOLDA
Sandia National Laboratories, Livermore, CA. Corresponding author: tgkolda@sandia.gov

AND

ALI PINAR
Sandia National Laboratories, Livermore, CA

[Received on 24 June 2016]

Network science is a powerful tool for analyzing complex systems in fields ranging from sociology to engineering to biology. This paper is focused on generative models of *bipartite graphs*, also known as *two-way graphs*. We propose two generative models that can be easily tuned to reproduce the characteristics of real-world networks, not just qualitatively, but quantitatively. The measurements we consider are the degree distributions and the bipartite clustering coefficient, which we refer to as the metamorphosis coefficient. We define edge, node, and degreewise metamorphosis coefficients, enabling a more detailed understand of the bipartite community structure. Our proposed bipartite Chung-Lu model is able to reproduce real-world degree distributions, and our proposed bipartite "BTER" model 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, metamorphosis coefficient, bipartite clustering coefficient, complex networks

1. Introduction

Network science is a powerful tool for analyzing 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, varying the characteristics of the graph to, for instance, test a graph algorithm under different scenarios. Ultimately, we hope to *understand* the underlying nature of complex systems, and modeling them mathematically is a way to test our understanding.

This paper is focused on generative models of *bipartite graphs*, also known as *two-way graphs*. Many real-world systems are naturally expressed as a bipartite graph, which is a graph whose vertices are divided into two partitions, U and V, such that edges, E, only connect vertices across the two partitions, i.e.,

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

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 rep-

resented 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 [6, 8, 9, 13, 21, 23] and references therein. An example bipartite graph is shown in Fig. 1.

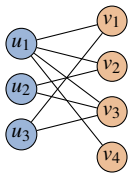


FIG. 1: Bipartite graph

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 are the degree distributions and the bipartite analog of the clustering coefficient. Introduced by Watts and Strogatz [25], the clustering coefficient of a one-way graph 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 [21] propose a bipartite clustering coefficient that is the probability that a bipartite three-path (or *caterpillar*[1]) participates in a bipartite four-cycle (or *butterfly*[2]), i.e.,

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

We call (1.1) the *metamorphosis coefficient*. 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. 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. In this paper, we extend (1.1) to define edge, node, and degreewise metamorphosis coefficients, enabling a more detailed understanding of the bipartite community structure.

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) [1, 3, 4], also known as the configuration model, and reproduces the degree distributions. 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

[1]Our usage of the term caterpillar is not to be confused with "caterpillar trees", which are trees in which all vertices are distance one from a central path.

[2]The term "butterfly" or "bowtie graph" has also been previously used to describe the graph on 5 vertices consisting of two triangles sharing a single vertex.

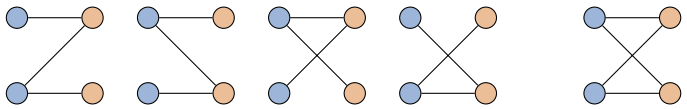


FIG. 2: A butterfly contains four distinct caterpillars

Erdős-Rényi (BTER) [11, 22] 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.

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 to 5. **CondMat** represents an author-paper network from arXiv preprints in condensed matter physics from 1995–99 [18]; this has mostly been used in the context of the coauthorship graph, but here we consider the underlying data [20]. The majority of authors have only a single 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 [12, 13], as collected from the Internet Movie Database. The busiest actor was in 294 movies; conversely, the largest production had 646 actors. **Flickr** [16, 17] 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** [5, 10] 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 (orange). This dataset apparently excludes critics with less than 20 reviews. The **MillionSong** [2, 14, 26] dataset connects users and the songs they played. The dataset only includes listeners of ten 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 [12, 13] 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** [16, 17] 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 five persons in that category. The largest group has over 1M members.

3. Notation

We set up the notation for one-way and two-way graphs. 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 for j and

Table 1: Real-world bipartite graphs

Name	Partition 1	Partition 2	Edges
CondMat [18, 20]	16,726 authors	22,016 papers	58,595
IMDB [12, 13]	127,823 actors	383,640 movies	1,470,418
Flickr [16, 17]	1,728,701 users	103,648 groups	8,545,307
MovieLens [5, 10]	65,133 movies	71,567 critics	10,000,054
MillionSong [2, 14, 26]	1,019,318 users	384,546 songs	48,373,586
Peer2Peer [12, 13]	1,986,588 peers	5,380,546 files	55,829,392
LiveJournal [16, 17]	5,284,451 users	7,489,296 groups	112,307,385

partition two. For instance, if $i = j = 2$ in the two-way case, they are referring to two distinct vertices.

Table 2: Notation

One-Way Graph $G = (V, E)$	**Two-Way Graph** $G = (U, V, E)$
Vertices : V	Vtx. Partition 1 : U
	Vtx. Partition 2 : V
Edges : $E \subseteq V \otimes V$	Edges : $E \subseteq U \otimes V$
# Vertices : $n = \lvert V \rvert$	# Vertices in U : $n^u = \lvert U \rvert$
	# Vertices in V : $n^v = \lvert V \rvert$
# Edges : $m = \lvert E \rvert$	# Edges : $m = \lvert E \rvert$
# Wedges : $n^{\wedge}$	# Caterpillars : $n^{\boxminus}$
# Triangles : $n^{\triangle}$	# Butterflies : $n^{\boxtimes}$
Clust. Coeff. : $c = 3n^{\triangle}/n^{\wedge}$	Meta. Coeff. : $c = 4n^{\boxtimes}/n^{\boxminus}$
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 : $d_i = \lvert \{ j \in V \mid (i, j) \in E \} \rvert$	Degree of i : $d_i^u = \lvert \{ j \in V \mid (i, j) \in E \} \rvert$
	Degree of j : $d_j^v = \lvert \{ i \in U \mid (i, j) \in E \} \rvert$

4. Fast Bipartite Chung-Lu Model

We adapt the Chung-Lu generative model [1, 3, 4] to bipartite graphs and demonstrate its ability to reproduce bipartite degree distributions. We follow the notation described in Section 3.

4.1 *Chung-Lu for One-way Graphs*

Consider a one-way graph $G = (V, E)$. The Chung-Lu (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 $d_i \leqslant \sqrt{2m}$ 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, i.e., the number of edges $m = O(n)$. For this reason, we favor "fast" CL that flips only $2m$ coins [7, 11]. We explain the fast method below in the context of two-way graphs.

4.2 *Chung-Lu 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 $d_i^u \leqslant \sqrt{m}$ for all i and $d_j^v \leqslant \sqrt{m}$ for all j. A naïve implementation of bipartite Chung-Lu 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

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

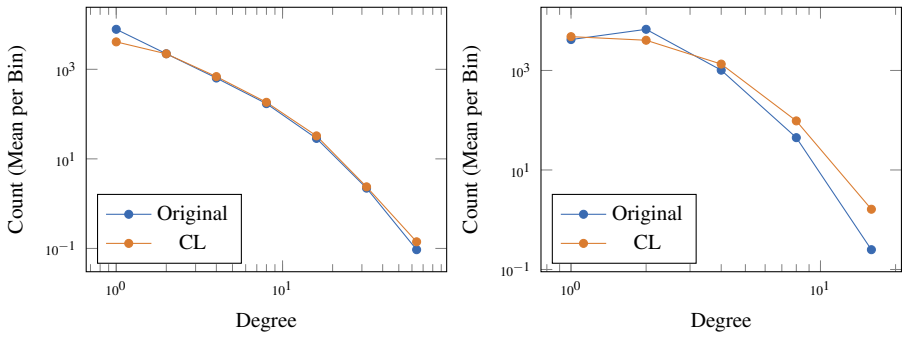
which is the same as (4.1). Although both implementations of CL yield identical expected degrees, a key distinction is that multiple edges between the same pair of vertices are possible in fast CL. In practice, for large graphs with heavy-tailed degree distributions, this rarely presents a problem. The fast CL algorithm is described in Algorithm 1. Another viewpoint on CL is the configuration model [9]: 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 two.

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

4.3 *Experimental Results*

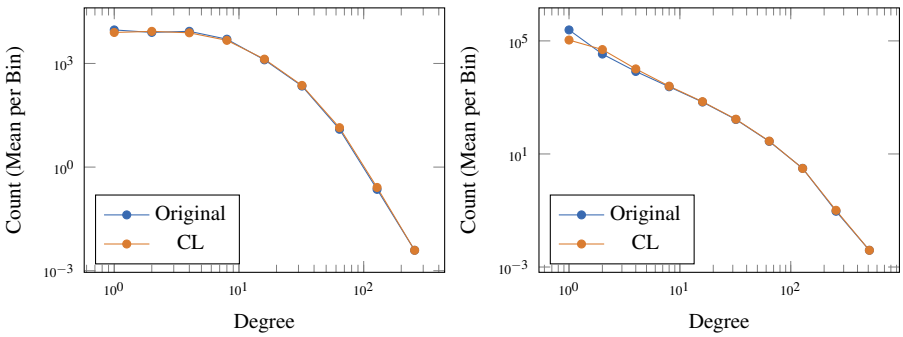
We generate random graphs using CL and the degree distributions of the graphs described in Section 2. The degree distributions are shown in Figs. 3 to 5. The degree distribution of the original graph is



(a) Mode 1: CondMat (Authors/Papers) (b) Mode 2: CondMat (Authors/Papers)

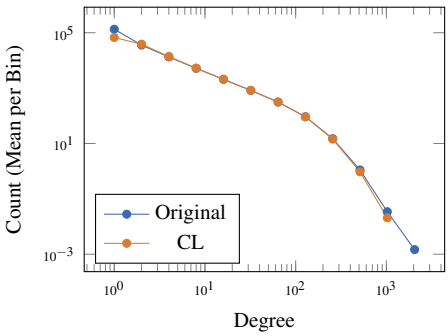
FIG. 3: Mode 2: CondMat (Authors/Papers)

shown in blue, and the degree distribution of the graph generated by CL is shown in orange. These are *binned* degree distributions, as advocated in [15]. 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, including zero values, so the y-coordinate can be less than one. An illustration of binned and raw data is presented in Appendix A. Overall, the degree distributions are very close, especially for IMDB (Figs. 4a and 4b), Flickr (Figs. 4c and 4d), and LiveJournal (Figs. 5e and 5f). For CondMat (Figs. 3a and 3b), the distribution of degrees on the author nodes is a close match, but there is some trouble matching the paper degree distribution. It 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 (Figs. 4e and 4f), the model generates a few "critic" nodes of degree less than 20, even though no nodes exist in the true degree distribution. A similar phenomena 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

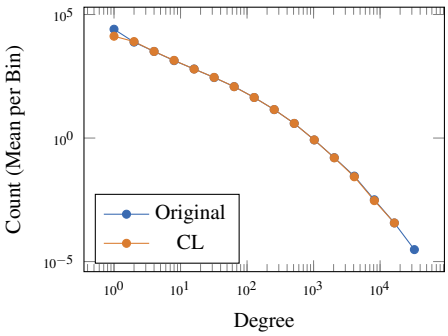


(a) Mode 1: IMDB (Actors/Movies)

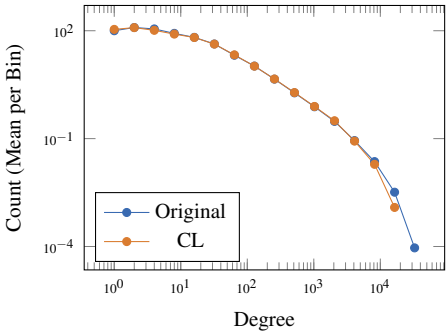
(b) Mode 2: IMDB (Actors/Movies)



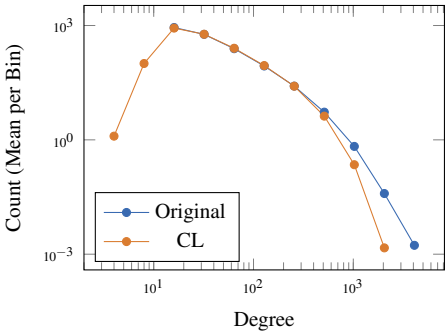
(c) Mode 1: Flickr (Users/Groups)



(d) Mode 2: Flickr (Users/Groups)

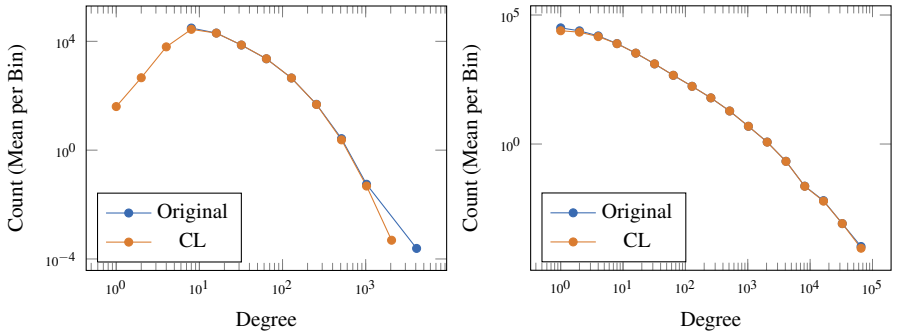


(e) Mode 1: MovieLens (Movies/Critics)



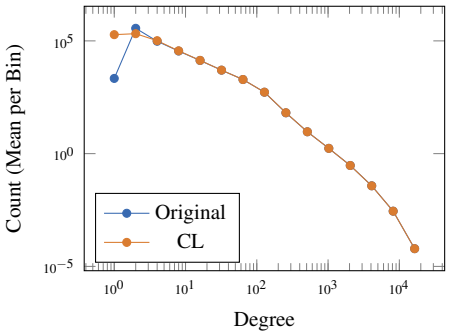
(f) Mode 2: MovieLens (Movies/Critics)

FIG. 4: Mode 2: MovieLens (Movies/Critics)

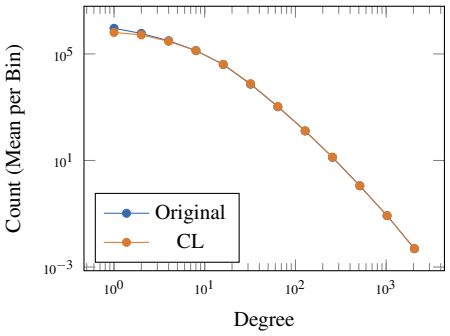


(a) Mode 1: MillionSong (Users/Songs)

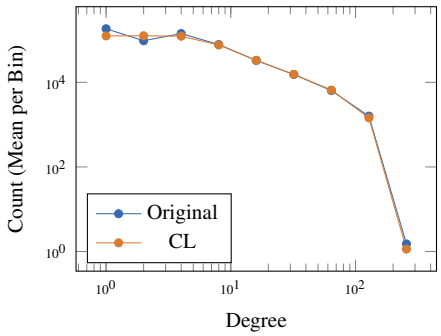
(b) Mode 2: MillionSong (Users/Songs)



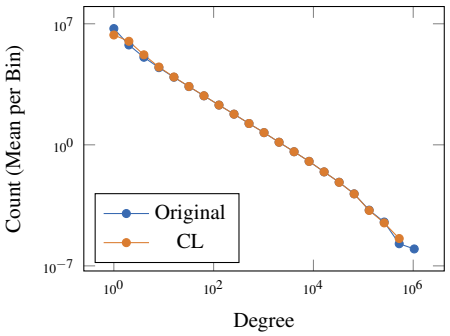
(c) Mode 1: Peer2Peer (Peers/Files)



(d) Mode 2: Peer2Peer (Peers/Files)



(e) Mode 1: LiveJournal (Users/Groups)



(f) Mode 2: LiveJournal (Users/Groups)

FIG. 5: Mode 2: LiveJournal (Users/Groups)

(Figs. 5c and 5d), the number of degree-1 peer nodes is underestimated, for reasons described in [7].

4.4 *Shortcomings of CL*

Overall, if we provide the degree distributions of real-world graph, the CL model and its implementation for bipartite graphs, FBCL, can generate a random graph, whose degree distribution closely matches the degree distribution of the original graph. However, we observe that the graphs generated by CL have very few butterflies, typically many fewer than the original graphs. Table 3 shows that the number of butterflies from 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 FBCL generated graphs on the other hand, is closer, so the metamorphosis of the original graph is much higher than for the CL graph with low butterfly counts. This indicates that the CL model is omitting some important structure.

Note that this structure what underlies the cohesive 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. And without the butterflies a graph will not have a community structure. This is our motivation for considering a more complex model in the next section.

Table 3: Properties of the original and the BCL graphs.

Graph	Size		Edges	Cats.	Buts.	Meta.
	n^u	n^v	m	$n^{\boxtimes}$	$n^{\boxtimes}$	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	2.57e12	3.53e10	5.49e−2
Flickr-CL	1.73e6	1.04e5	8.39e6	2.20e12	1.52e10	2.78e−2
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
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
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
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

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 [11, 22]. 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, Magnien, and Vecchio [13] and Opsahl [19].

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

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

We know the number of caterpillars immediately from the degrees of i and j, i.e.,

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

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:

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

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, i.e.,

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

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 Figs. 6c and 6d. 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. An illustration of raw and binned metamorphosis coefficients is presented in Appendix A. 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 an impossibility.

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 is 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), i.e.,

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

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

$$c_i^u = \frac{1}{d_i^u} \sum_{(i,j)\in \hat{E}} \frac{n^{\boxtimes}_{(i,j)}}{n^{\boxplus}_{(i,j)}} = \frac{1}{d_i^u} \sum_{(i,j)\in \hat{E}} \frac{n^{\boxtimes}_{(i,j)}}{(d_i^u - 1)(d_j^v - 1)}. \tag{5.6}$$

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

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

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,

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

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). 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:

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

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

5.3 *Bipartite BTER algorithm*

The affinity blocks are constructed as explained in Section 5.2, and then the *excess degree* (i.e., edges not used in constructing the affinity blocks) is connected via a fact bipartite CL procedure. The algorithm is listed in Algorithm 2.

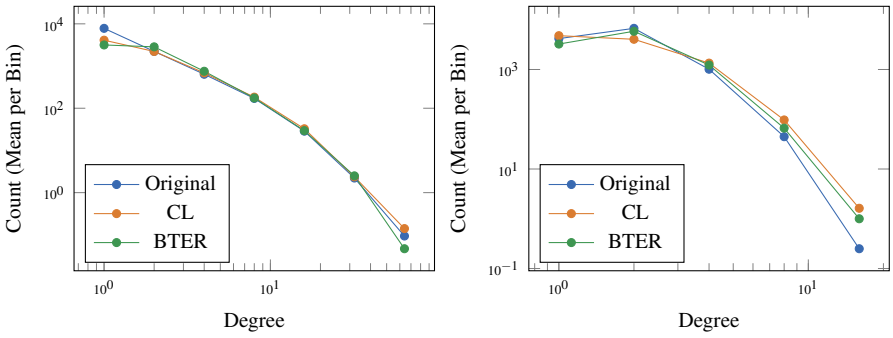
5.4 *Experimental Results*

We generate bipartite BTER (biBTER) 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 Figs. 6a and 6b, show little difference between biBTER and CL; both are good matches to the original degree distribution. The degreewise metamorphosis coefficients are shown in Figs. 6c and 6d. Although there is not a perfect match between biBTER 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 biBTER. The number of butterflies and metamorphosis coefficients are significantly

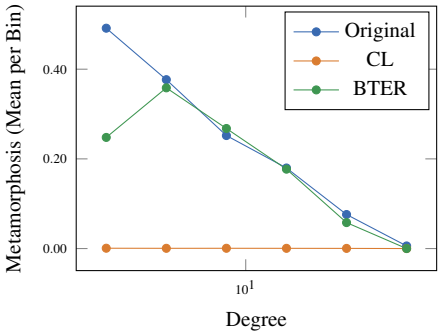
Algorithm 2 Bipartite BTER

procedure BIBTER($\{d_i^u\}, \{d_j^v\}, \{c_d^u\}, \{c_d^v\}$)
We assume that degrees are sorted in increasing order
  $m \leftarrow \sum_i d_i^u$
  $E \leftarrow \emptyset$
  $\{e_i^u\} \leftarrow \{d_i^u\}, \{e_j^v\} \leftarrow \{d_j^v\}$ ▷ Excess degree initialization
  $i \leftarrow \min\{i \mid d_i^u > 1\}, j \leftarrow \min\{j \mid d_j^v > 1\}$
  repeat ▷ Create affinity blocks until nodes exhausted
    $\hat{d}^u \leftarrow d_i^u, \hat{d}^v \leftarrow d_j^v$
    $\hat{c}^u \leftarrow c_{\hat{d}^u}, \hat{c}^v \leftarrow c_{\hat{d}^v}$
    if $\hat{c}^u/\hat{c}^v \geqslant 1$ **then**
      $\hat{n}^u \leftarrow \hat{d}^v, \hat{n}^v \leftarrow \texttt{round}\left(\frac{\hat{c}^u}{\hat{c}^v}\hat{d}^u\right), \rho \leftarrow \left(\frac{(\hat{d}^u-1)(\hat{c}^v)^2}{\hat{c}^u\hat{d}^u-\hat{c}^v}\right)^{1/4}$
    else
      $\hat{n}^v \leftarrow \hat{d}^u, \hat{n}^u \leftarrow \texttt{round}\left(\frac{\hat{c}^v}{\hat{c}^u}\hat{d}^v\right), \rho \leftarrow \left(\frac{(\hat{d}^v-1)(\hat{c}^u)^2}{\hat{c}^v\hat{d}^v-\hat{c}^u}\right)^{1/4}$
    end if
    if $i+\hat{n}^u-1 \leqslant n^u$ and $j+\hat{n}^v-1 \leqslant n^v$ **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**
            $E \leftarrow E \cup (\hat{i}, \hat{j})$
            $e_{\hat{i}}^u = \max\{0, e_{\hat{i}}^u - 1\}, e_{\hat{j}}^v = \max\{0, e_{\hat{j}}^v - 1\}$ ▷ 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 $j > n^v$
  $E \leftarrow E \cup \text{FBCL}(\{e_i^u\}, \{e_j^v\})$
end procedure

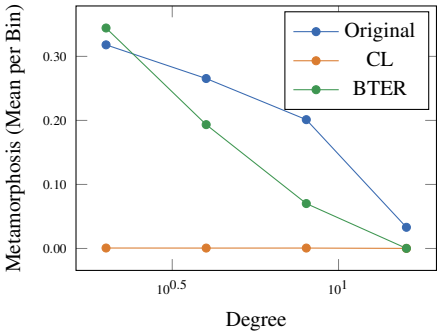


(a) Mode 1: CondMat (Authors/Papers)

(b) Mode 2: CondMat (Authors/Papers)



(c) Mode 1: CondMat (Authors/Papers)



(d) 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 biBTER.

Table 4: Metamorphosis coefficients for original versus BCL graphs.

Graph	Size		Edges	Cats.	Buts.	Meta.
	n^u	n^v	m	$n^{\triangle}$	$n^{\boxtimes}$	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	6.01e4	2.33e6	1.16e5	1.99e−1
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	6.70e6	1.99e−2
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	4.26e10	7.21e−2
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
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

improved as compared to CL. We see that CondMat has 70,000 butterflies, CL produces less than 400 butterflies, but biBTER produces 120,000 butterflies. The biBTER number is a slight overestimate, but overall much better than CL.

The degreewise metamorphosis coefficients for the remaining graphs are shown in Figs. 7 and 8. Even in cases where CL's overall numbers are close to the original graph as shown in Table 4, the degreewise metamorphosis coefficients are low. biBTER corrects this problem and gets metamorphosis coefficients that are much closer to what we have seen previously. In some cases like MovieLens (Fig. 7e), the CL metamorphosis coefficients are nonzero, which is the result of the overall high density of the graph. Indeed, it has been proven (see Lemma 1 in [24]) 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 biBTER. MovieLens also have an unusual profile in Mode 2 (see Fig. 7f), which is really just an artifact of the data collection. Nevertheless, biBTER is able to obtain a reasonable approximation. For completeness, the degree distributions for CL and biBTER as compared to the original graphs are shown in Figs. 9 and 10. There is little difference between biBTER and CL in terms of the degree distribution.

6. Conclusions

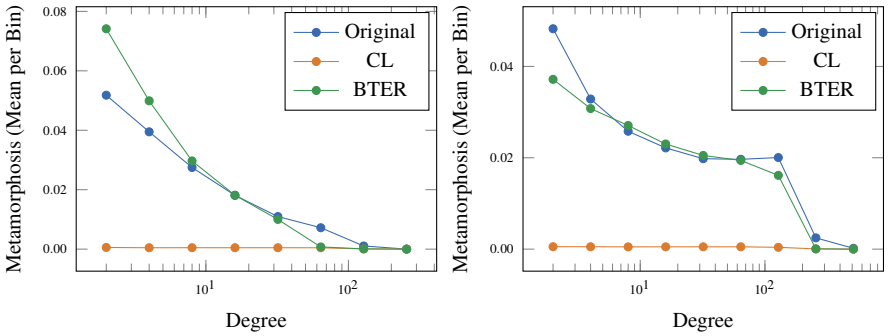
We have considered the problem of how to generate realistic bipartite graph models the reproduce the characteristics of large real-world networks. Our first proposed model, bipartite CL, accurately reproduces the degree distribution. Our second proposed model, bipartite BTER, goes further to capture the degreewise metamorphosis coefficients. High coefficients are indicative of a relatively large number of butterflies indicating cohesion or community structure, which is rare is sparse graphs unless there is 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 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. 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. These are but a few topics for future investigation.

A. Binned Distributions

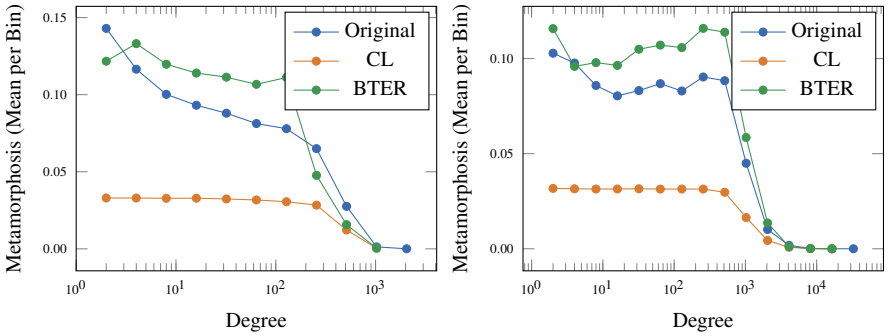
As mentioned in the main text, we use *binned* degree distributions, as advocated in [15]. We use powers of two for the bins, so the x-coordinate 2^k corresponds to the bin from $[2^k, 2^{k+1})$. The y-coordinate is the average value for that bin, including zero values, so the y-coordinate can be less than one.

A comparison of binned and raw degree distributions for CondMat is shown in Fig. A.11. This is a



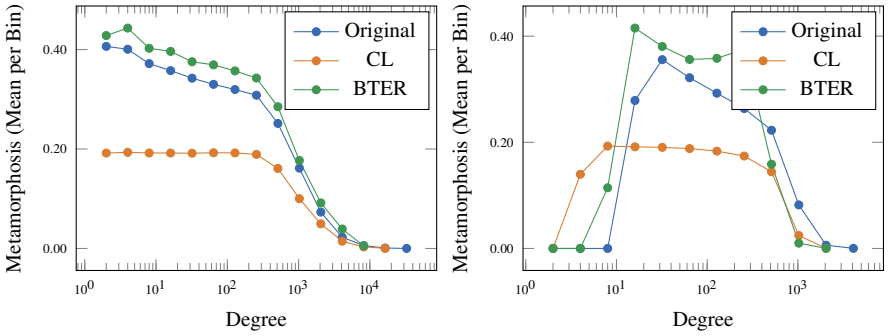
(a) Mode 1: IMDB (Actors/Movies)

(b) Mode 2: IMDB (Actors/Movies)



(c) Mode 1: Flickr (Users/Groups)

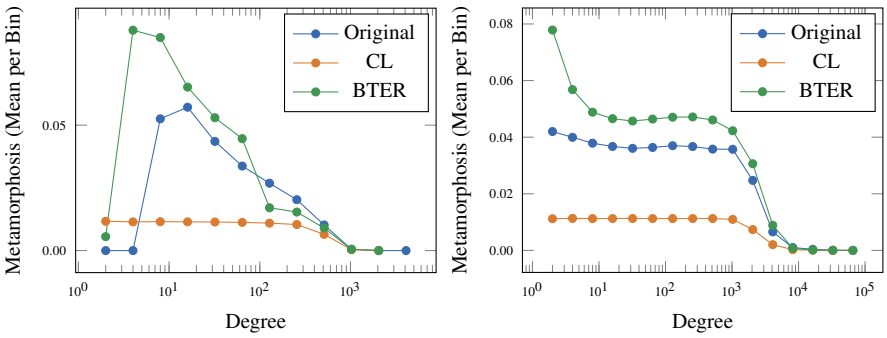
(d) Mode 2: Flickr (Users/Groups)



(e) Mode 1: MovieLens (Movies/Critics)

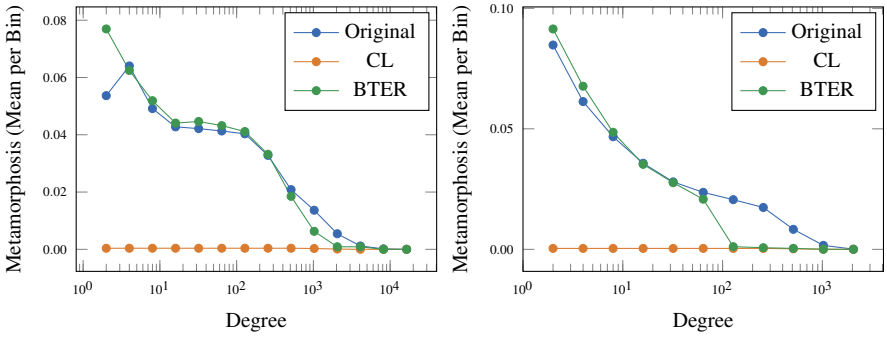
(f) Mode 2: MovieLens (Movies/Critics)

FIG. 7: Mode 2: MovieLens (Movies/Critics)



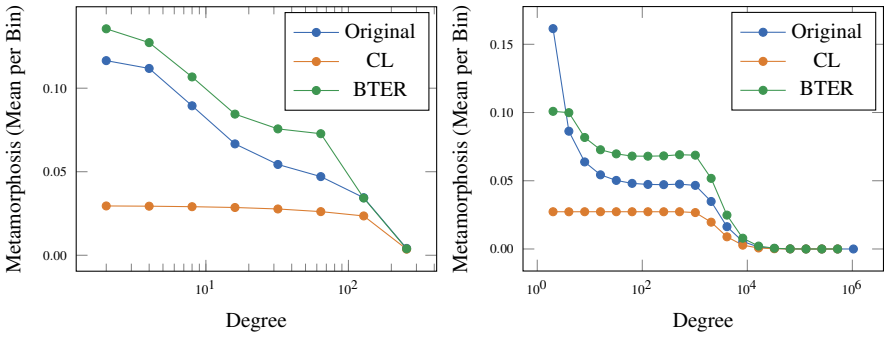
(a) Mode 1: MillionSong (Users/Songs)

(b) Mode 2: MillionSong (Users/Songs)



(c) Mode 1: Peer2Peer (Peers/Files)

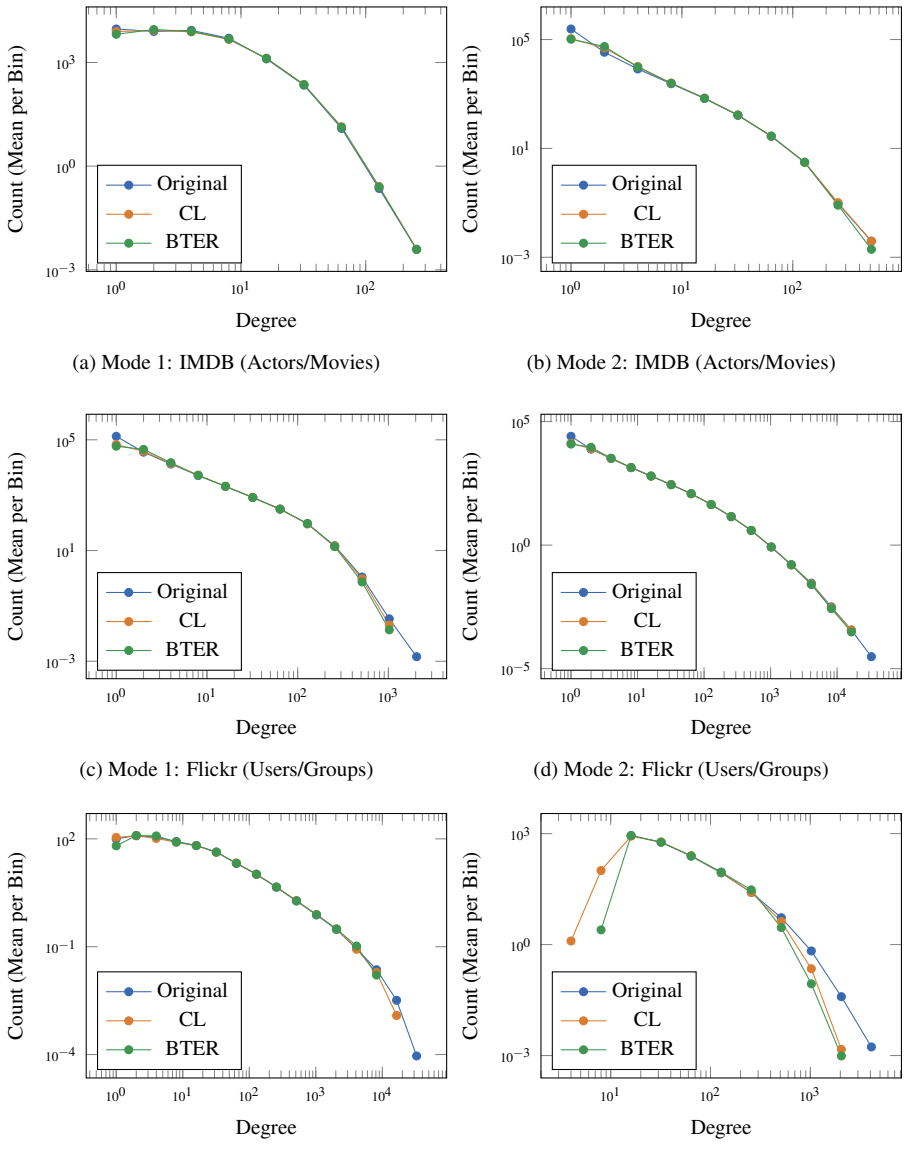
(d) Mode 2: Peer2Peer (Peers/Files)



(e) Mode 1: LiveJournal (Users/Groups)

(f) Mode 2: LiveJournal (Users/Groups)

FIG. 8: Mode 2: LiveJournal (Users/Groups)



(a) Mode 1: IMDB (Actors/Movies)

(b) Mode 2: IMDB (Actors/Movies)

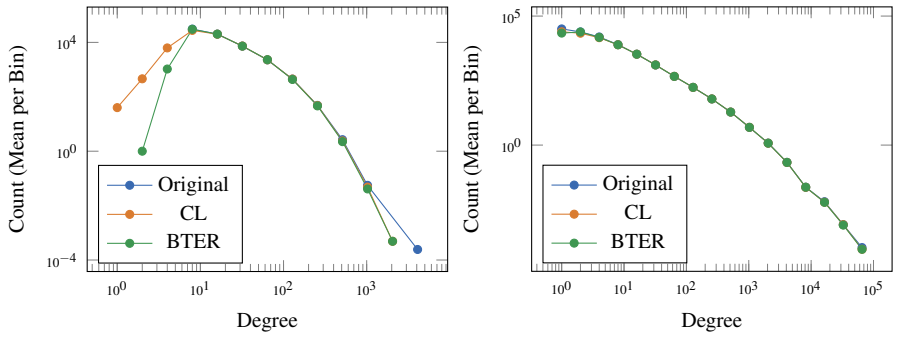
(c) Mode 1: Flickr (Users/Groups)

(d) Mode 2: Flickr (Users/Groups)

(e) Mode 1: MovieLens (Movies/Critics)

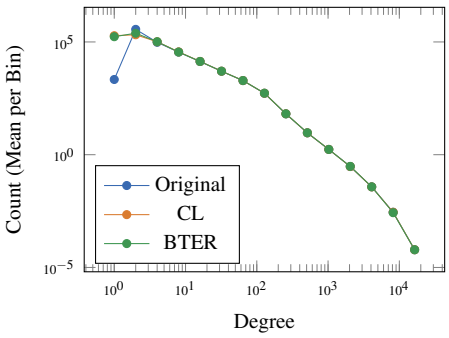
(f) Mode 2: MovieLens (Movies/Critics)

FIG. 9: Degree distributions illustrating the original (blue), fast BCL (orange), and bipartite BTER (green). The data is log-binned.

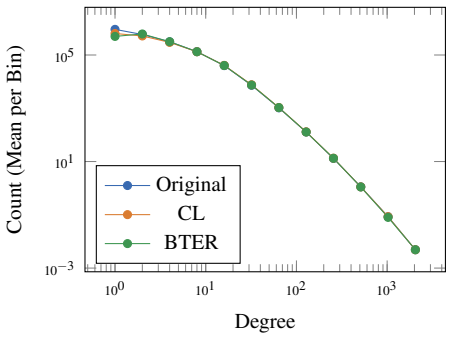


(a) Mode 1: MillionSong (Users/Songs)

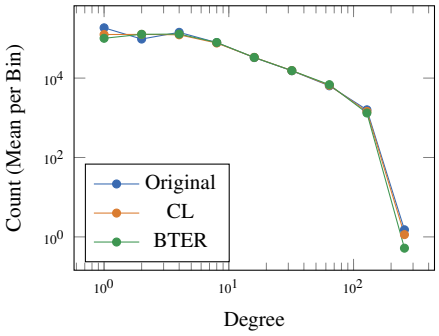
(b) Mode 2: MillionSong (Users/Songs)



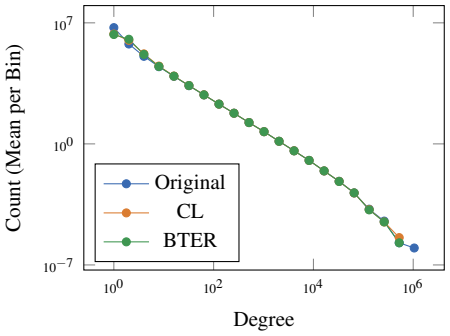
(c) Mode 1: Peer2Peer (Peers/Files)



(d) Mode 2: Peer2Peer (Peers/Files)



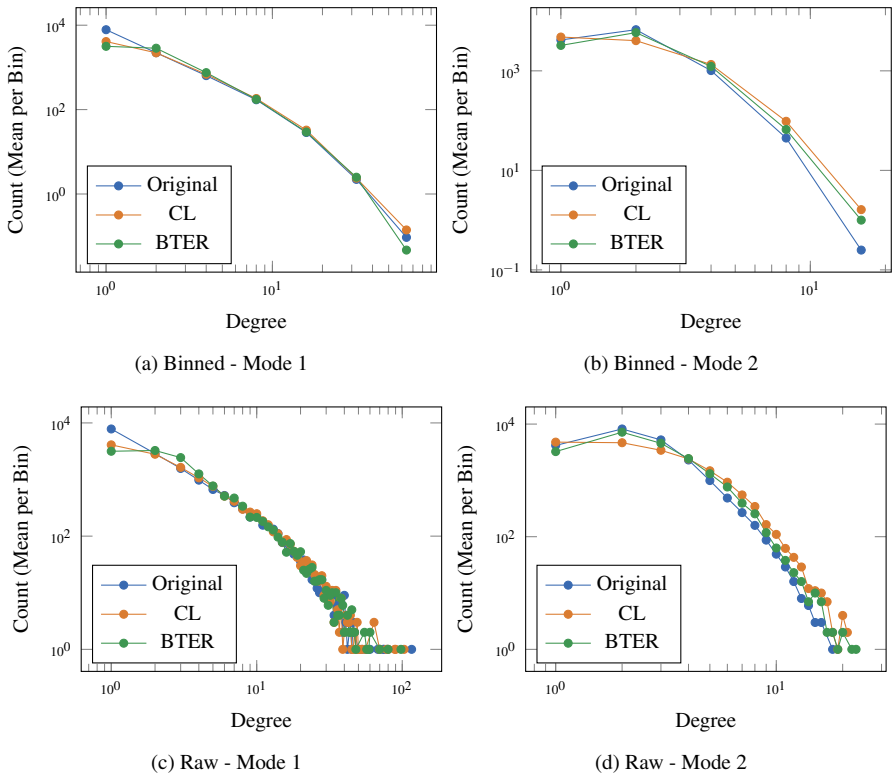
(e) Mode 1: LiveJournal (Users/Groups)



(f) Mode 2: LiveJournal (Users/Groups)

FIG. 10: Degree distributions illustrating the original (blue), fast BCL (orange), and bipartite BTER (green). The data is log-binned.

relatively small graph, but here is becomes clear how difficult it is to compare degree distributions for the raw data. Note that neither plot shows zeros explicitly.



(a) Binned - Mode 1

(b) Binned - Mode 2

(c) Raw - Mode 1

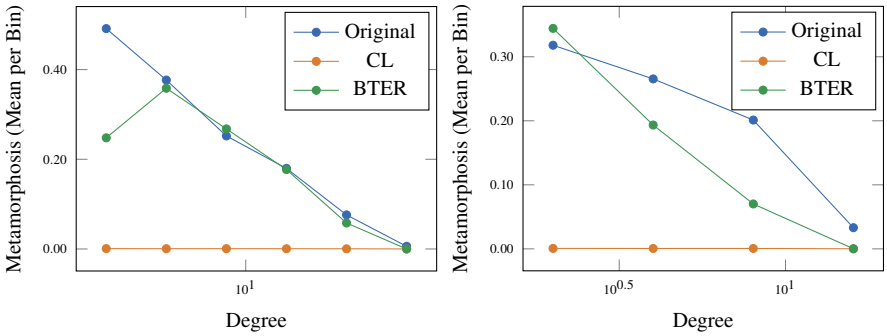
(d) Raw - Mode 2

FIG. A.11: Binned versus raw degree distributions (CondMat)

Likewise, the binned and raw metamorphosis coefficients are illustrated in Fig. A.12. Here, we do not plot the zeros (though we do plot small values) in the raw data. The data has outliers, mostly due to the small number of vertices of each degree, especially for mode 2. The binning has the effect of smoothing the data.

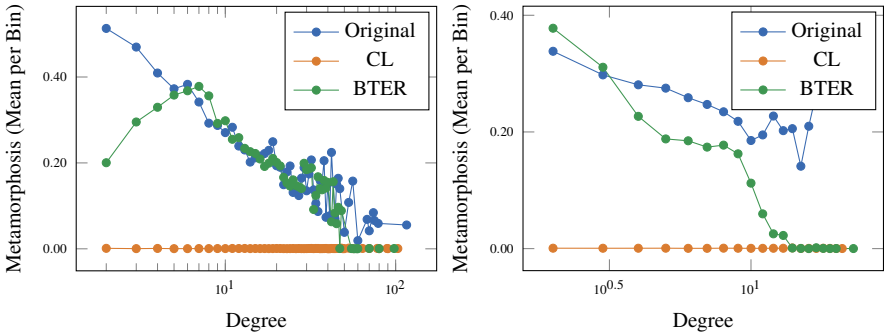
Funding

This work was supported in part by the Defense Advanced Research Project Agency (DARPA). Sandia National Laboratories is a multi-program 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.



(a) Binned - Mode 1

(b) Binned - Mode 2



(c) Raw - Mode 1

(d) Raw - Mode 2

FIG. A.12: Binned versus raw metamorphosis coefficients (CondMat)

Acknowledgments

We thank our colleagues for helpful discussions in the course of this work, especially Gray Ballard, Kenny Chowdhary, Danny Dunlavy, Kevin Matulef, and Michael Wolf.

REFERENCES

1. Aiello, W., Chung, F. & Lu, L. (2001) A random graph model for power law graphs. *Experimental Mathematics*, **10**, 53–66.
2. 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*. University of Miami.
3. Chung, F. & Lu, L. (2002a) The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences*, **99**(25), 15879–15882.
4. Chung, F. & Lu, L. (2002b) Connected components in random graphs with given degree sequences. *Annals of Combinatorics*, **6**, 125–145.
5. Davies, R. (2009) MovieLens 10M. GroupLens, Department of Computer Science and Engineering, University of Minnesota.
6. 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*, pages 269–274, New York, NY, USA. ACM.
7. 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*, pages 23–30.
8. Guillaume, J.-L. & Latapy, M. (2004) Bipartite structure of all complex networks. *Information Processing Letters*, **90**(5), 215221.
9. Guillaume, J.-L. & Latapy, M. (2006) Bipartite Graphs as Models of Complex Networks. *Physica A*, **371**, 795–813.
10. Harper, F. M. & Konstan, J. A. (2015) The MovieLens Datasets. *ACM Transactions on Interactive Intelligent SystemsMar*, **5**(4), 1–19.
11. Kolda, T. G., Pinar, A., Plantenga, T. & Seshadhri, C. (2014) A Scalable Generative Graph Model with Community Structure. *SIAM Journal on Scientific Computing*, **36**(5), C424–C452.
12. Latapy, M. (2006) Bipartite Graph Tools, Version 1.0. .
13. Latapy, M., Magnien, C. & Vecchio, N. D. (2008) Basic Notions for the Analysis of Large Two-mode Networks. *Social Networks*, **30**(1), 31–48.
14. 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*, pages 909–916.
15. Milojević, S. (2010) Power Law Distributions in Information Science: Making the Case for Logarithmic Binning. *Journal of the American Society for Information Science and Technology*, **61**(12), 2417–2425.
16. Mislove, A. (2015) Personal Communication. .
17. 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*, pages 29–42, San Diego, CA.
18. Newman, M. E. J. (2001) The Structure of Scientific Collaboration Networks. *Proceedings of the National Academy of Sciences*, **98**(2), 404–409.
19. Opsahl, T. (2013) Triadic Closure in Two-mode Networks: Redefining the Global and Local Clustering Coefficients. *Social Networks*, **35**(2), 159–167.
20. Opsahl, T. (2015) Tnet Datasets. .
21. Robins, G. & Alexander, M. (2004) Small Worlds Among Interlocking Directors: Network Structure and Distance in Bipartite Graphs. *Computational & Mathematical Organization Theory*, **10**(1), 69–94.
22. Seshadhri, C., Kolda, T. G. & Pinar, A. (2012) Community Structure and Scale-free Collections of Erdős-Rényi Graphs. *Physical Review E*, **85**(5).

23. Sun, J., Qu, H., Chakrabarti, D. & Faloutsos, C. (2005) Relevance Search and Anomaly Detection in Bipartite Graphs. *SIGKDD Explorations Newsletter*, **7**(2), 48–55.
24. Tait, M. & Verstraëte, J. (2016) On Sets of Integers with Restrictions on Their Products. *European Journal of Combinatorics*, **51**, 268–274.
25. Watts, D. & Strogatz, S. (1998) Collective dynamics of 'small-world' networks. *Nature*, **393**, 440–442.
26. The Echo Nest Taste Profile Subset, the Official User Data Collection for the Million Song Dataset. `http://labrosa.ee.columbia.edu/millionsong/tasteprofile`.