Soc. Netw. Anal. Min. (2011) 1:187–218
DOI 10.1007/s13278-011-0021-0

ORIGINAL ARTICLE

A systematic approach to the one-mode projection of bipartite graphs

Katharina Anna Zweig · Michael Kaufmann

Received: 1 November 2010 / Accepted: 22 February 2011 / Published online: 6 April 2011

Abstract Bipartite graphs are common in many complex systems as they describe a relationship between two different kinds of actors, e.g., genes and proteins, metabolites and enzymes, authors and articles, or products and consumers. A common approach to analyze them is to build a graph between the nodes on one side depending on their relationships with nodes on the other side; this so-called *one-mode projection* is a crucial step for all further analysis but a systematic approach to it was lacking so far. Here, we present a systematic approach that evaluates the significance of the *co-occurrence* for each pair of nodes v, w, i.e., the number of common neighbors of v and w. It turns out that this can be seen as a special case of evaluating the *interestingness* of an *association rule* in data mining. Based on this connection we show that classic interestingness measures in data mining cannot be applied to evaluate most real-world product-consumer relationship data. We thus introduce generalized interestingness measures for both, one-mode projections of bipartite graphs and data mining and show their robustness and stability by example. We also provide theoretical results that show that the old method cannot even be used as an approximative method. In a last step we show that the new interestingness measures show stable and significant results that result in attractive one-mode projections of bipartite graphs.

K. A. Zweig (✉)
Interdisciplinary Center for Scientific Computing,
University of Heidelberg, Speyerer Straße 6,
69115 Heidelberg, Germany
e-mail: katharina.zweig@iwr.uni-heidelberg.de

M. Kaufmann
Wilhelm-Schickard-Institute, University of Tübingen,
Sand 14, 72072 Tübingen, Germany
e-mail: mk@informatik.uni-tuebingen.de

Keywords Bipartite graphs · One-mode projection · Association rules · Interestingness measures

1 Introduction

Many relationships in complex systems are best represented by a bipartite graph: metabolites that are transformed by enzymes (Ravasz et al. 2002), scientists that co-author a paper (Newman 2001a, b), finches and their island habitats (Cobb and Chen 2003), genes and the diseases with which they are associated (Goh et al. 2007), or products that are bought by consumers (Gionis et al. 2007). Since most tools in network analysis are focused on general graphs, it is a common approach to analyzing bipartite networks to project them onto one of their sides. In this so-called *one-mode projection* the nodes on one of the sides are connected with each other according to their connection pattern to nodes on the other side; the nodes of the other side are discarded (Wasserman and Faust 1999, Chap 8). With this approach, metabolites are connected by an edge if they are transformed into each other by an enzyme, scientists are connected if they have written papers together, and genes are connected if they cause the same disease. Vice versa, enzymes are connected if they transform the same metabolites, papers are connected if they share an author, and two different disease are connected if they are caused by the same gene.

Zhou et al. have already stated that a reasonable one-mode projection of product-consumer networks can be used for deriving recommendations (Zhou et al. 2007). Classic recommendation systems are often built on so-called *association rules*, i.e., on rules like ‘if product A and B are bought, then product C, D, and E are also often bought’. If such an association rule is true for many consumers, then it is natural to recommend products C, D, and

E to any person who has bought A and B. We see an even deeper connection between the two research questions of finding interesting association rules and of finding a reasonable one-mode projection. In this article we will show that association rules between single products, i.e., of the type 'if A then B', can be used to build a reasonable one-mode projection. Vice versa, the resulting graph can then be used to derive candidates for more complex association rules, as, for example, proposed by Raeder and Chawla (2011). We also show on two examples that the classic interestingness measures for association rules need to be generalized with respect to new null-models.

We show the usefulness of one of the newly generalized interestingness measures, the $leverage_{\text{FDSM}}$, as the basis for a one-mode projection on a data set of film ratings. The $leverage_{\text{FDSM}}$ assigns a real value to all pairs of films, and thus, for each film all other films can be ranked according to their $leverage_{\text{FDSM}}$ value. To evaluate whether the topmost ranked films are really suitable neighbors in a one-mode projection of the graph, a subset of all films is extracted that contains parts of series. For almost all films in this set, our automatic and otherwise oblivious algorithm is able to identify the other parts of the same series among the top ten ranks, using only the rating pattern of a subset of 20,000 users. This implies that a reasonable one-mode projection can be achieved by connecting each vertex with its topmost ranked other films, as we will discuss in this article. The article thus comprises results for two different communities, network analysis and data mining.

The article is based on a shorter version by one of the authors (Zweig 2010), extended by additional theoretical and empirical results, and it is organized as follows: Sect. 2 gives the necessary definitions, and Sect. 3 describes the state of the art and the relationship between one-mode projections of bipartite graphs and the evaluation of association rules. Section 4 introduces the new method to build a sparse one-mode projection of bipartite graphs, followed by theoretical results on the new and the classic method in Sect. 5. In Sect. 6 we show experimental results on a large film rating-consumer data set.[1] In Sect. 7.3 we summarize the implications for market basket analysis based on association rules. We conclude the article by open questions and a discussion in Sect. 7.

2 Definitions

The definitions are given within the context of product-consumer networks but they are applicable to any kind of bipartite graphs. Let $U = \{u_1, u_2,\ldots, u_r\}$ denote a set of *users* or customers, and $P = \{p_1, p_2,\ldots, p_l\}$ a set of *products*. Let $E \subseteq U \times P$ be a set of pairs of users and products u_x, p_i, denoting that user u_x has bought product p_i. Note that E is a set, not a multi–set, i.e., we assume that each user buys each product at most once. The sets U, P, E can also be represented by a bipartite graph $G = (U \cup P, E)$ where U is a set of r and P a set of l vertices that are connected by an edge iff $(u_x, p_i) \in E$ with $1 \leq x \leq r$ and $1 \leq i \leq l$. We will denote the vertex and the represented object by the same label as long as there is no ambiguity. Let then $m: = |E|$ denote the cardinality of this set; m is at the same time the number of edges in the bipartite graph. By $deg(u_x)$ we denote the *degree* of user u_x, i.e., the cardinality of the set of pairs in E that contain u_x. Analogously, $deg(p_i)$ is defined as the cardinality of the set of pairs in E that contain p_i. Note that $\forall\, u_x, deg(u_x) \leq l$ and $\forall\, p_j, deg(p_j) \leq r$. If the data is represented in a 0-1 table where products are in rows and users in columns, then $deg(p_i)$ is equal to the i-th row sum, and $deg(u_x)$ is equal to the x-th column sum.

The nodes on both sides of the graph are identified by an index from 1 to l and 1 to r, respectively. As the *left-hand degree sequence* L we define the sequence of degrees of nodes in P, as the *right-hand degree sequence* R we define the sequence of degrees of nodes in U, sorted by their respective degree. If only the degree sequences of a bipartite graph were known, then the best estimate of the probability that a user u_x has bought product p_i is given by $P(p_i) = deg(p_i)/r$. This estimate is also called the *support* of p_i, denoted by $supp(p_i)$. Analogously, the probability that any product p_i drawn uniformly at random was bought by user u_x is given by $P(u_x) = deg(u_x)/l$. Given a bipartite graph $G = (U \cup P, E)$, for any pair of products $p_i, p_j,\ i \neq j$ we define as their *co-occurrence*, denoted by $coocc_G(p_i, p_j)$, the cardinality of the set of users that bought both products, i.e., of all users u_x where (u_x, p_i) and $(u_x, p_j) \in E$. The *support*, denoted by $supp_G(p_i, p_j)$, is then defined as $coocc_G(p_i, p_j)/r$. The support can be interpreted as the probability that any user drawn uniformly at random has bought both products. Since we exclude self-loops, $coocc_G(p_i, p_i)$ is not defined. Note that we might omit the index if the graph G is clearly defined in the given context. The definitions can be generalized for *subsets of products* X, where $supp(X)$ is the fraction of users who bought **all** products in X and $coocc(X)$ is their absolute number.

By $\mathcal{G}(L,R)$ we denote the *set of all bipartite graphs* $G' = (U \cup P, E')$ that obey given degree sequences L, R. Note that we allow neither multi-edges nor self–loops. Given any two degree sequences L and R, it is easy to decide whether the set $\mathcal{G}(L,R)$ is empty or not (Brualdi 1980, 2006): Let $\hat{R}$ and $\hat{L}$ denote the non-increasingly

[1] This part reproduces parts of the conference version published with IEEE (Zweig 2010). Reproduced with permission.

sorted versions of R and L, i.e., their *monotone rearrangements*. Let R^* denote $\hat{R}$'s *conjugated vector*, defined as:

$$R_j^* = |\{1 \leq i \leq r | deg(u_i) \geq j\}|, \tag{1}$$

i.e., R^* contains at j the number of elements in R that are not smaller than j. Note that R^* is by definition non-decreasingly sorted. R^* is said to *majorize* vector $\hat{L}$ if

$$\sum_{i=0}^{k} \hat{l}_i \leq \sum_{j=0}^{k} r_j^*, \quad \forall k \leq l, \tag{2}$$

with equality for $k = l$. According to the Gale (1957)/Ryser (1963, pp. 63–65)/Ford and Fulkerson (1962, pp. 79–82) theorem, $\mathcal{G}(L, R)$ is non–empty iff R^* dominates $\hat{L}$.

With these definitions we will now portrait the connection between association rules, the significance of network motifs, and one-mode projections of bipartite graphs.

3 State of the art

3.1 Association rules

Some products are often bought together and their placement in a supermarket might be a crucial element for the market's success. Similarly, if many customers like film A and film B, it is reasonable to recommend film B to all customers who already watched and liked film A. The understanding of which products are most often bought together thus provides important information for recommendation systems and shop design. This information about products that are frequently bought together can be computed by a so-called *market basket analysis* where each basket and each product is represented by a node, and a basket is connected by an edge with all the products it contains. To understand which products are likely to be bought together, a market basket analysis tries to identify which subsets of products are *significantly* bought more often together than expected by pure chance. The absolute number of times subsets of products have been bought together, i.e., the number of their *co-occurrence* in market baskets, fails as a measure if some products are in general bought much more often than others. This is obviously the case for most real-world product-consumer relationships in which butter and bread are bought on a daily basis while TVs and cars are not, and where a few books and films become bestsellers while many never sell more than a few hundreds. Association rules are simple implications of the type $X \rightarrow Y$ where X and Y are subsets of products (with $X \cap Y = \emptyset$) which are extracted from a given data set (Agrawal et al. 1993; Hipp et al. 2000). Let $P(XY)$ denote the fraction of baskets that contain all products from both subsets X and Y. Any association rule can be assigned an *interestingness* that tries to capture how useful the rule is. Piatetsky-Shapiro proposed that a good interestingness measure F should fulfill the following requirements (Piatetsky-Shapiro 1991):

1. $F(X,Y) = 0$ if X and Y are statistically independent, that is, $P(XY) = P(X)P(Y)$.
2. Monotonicity:
   (a) $F(X, Y)$ increases monotonically with $P(XY)$ when $P(X)$ and $P(Y)$ remain the same;
   (b) $F(X, Y)$ decreases monotonically with $P(X)$ (or $P(Y)$) when $P(XY)$ and $P(Y)$ (or $P(X)$) remain the same.

In this article we will concentrate on two intuitive interestingness measures, called *leverage* (introduced in Piatetsky-Shapiro 1991) and *lift* (introduced in Brin et al. 1997). Let the association rule of interest be $X \rightarrow Y$. The *leverage* $lev(X, Y)$ of two subsets of products is designed to measure the *difference between the observed and the expected support* of X and Y (Piatetsky-Shapiro 1991). It is defined as

$$lev(X, Y) = supp(X, Y) - supp(X) \cdot supp(Y) \tag{3}$$

$$lev(X, Y) = \frac{coocc(X, Y)}{r} - \frac{deg(X) \cdot deg(Y)}{r^2}. \tag{4}$$

The *lift* describes the fraction between the observed and the expected support of X and Y and is defined as

$$lift(X, Y) = \frac{supp(X, Y)}{supp(X) \cdot supp(Y)} \tag{5}$$

$$lift(X, Y) = \frac{coccc(X, Y) \cdot r}{deg(X) \cdot deg(Y)}. \tag{6}$$

Both measures contain a term that describes the expected co-occurrence. The idea behind the expectation model is that—when two probabilities are independent—the probability to observe both events at the same time is described by the product of the single probabilities. Thus, under the assumption of independence, the probability (or frequency) with which we expect two subsets of products to be bought is given by the product of their respective probabilities. In summary, the underlying simple independence model assumes the following two equivalences to hold if the subsets of products X and Y are bought independently:

$$P(X, Y) = P(X)P(Y) \tag{7}$$

and

$$P(X \mid Y) = P(X, Y)/P(Y) = P(X). \tag{8}$$

Although the theory behind all of these interestingness measures is very clear and straightforward, it is known for long that the results are difficult to interpret. One of the first article on association rules analyzed the US census data (Brin et al. 1997). Brin et al. state that it is very difficult to quantify how well association rules (which they call implication rules) work. The authors state ironically, that the most *interesting* rules they found where: … five year olds don't work, unemployed resident's don't earn income from work, men don't give birth, and many other interesting facts.[2]

In their summary they write:

> Looking over the implication rules generated on census data was educational. First, it was educational because most of the rules themselves were not. The rules that came out on top were things that were obvious.

They propose to look for association rules with a medium interestingness. However, of those rules there were many (about 20,000). We will show in the following that a large number of seemingly interesting association rules might be based on the wrong underlying independence model and that a network analytic perspective on evaluating the interestingness of association rules relieves this problem to a large extent.

3.2 One-mode projection

Bipartite graphs are common in all kinds of complex systems and the co-actor network used in Watts' and Strogatz' seminal paper was actually a one-mode projection of the bipartite film-cast network (Watts and Strogatz 1998). They chose the most simple method by connecting any two actors if they were casted for at least one common film. In general, this method introduces many cliques in the resulting graph and does not differentiate between two nodes that share many neighbors and those that share only one. To relieve the latter, a simple weight on the edges can present the number of shared neighbors. This weight can then also be used to remove edges with a weight below some chosen threshold t. The simplest approach above is then equal to setting $t \equiv 1$.

A different approach focuses on how to treat nodes with very different degrees, e.g., articles with only two authors or articles with many authors. To address this problem, Newman suggested to assign a weight to all pairs of neighbors of v that is inversely proportional to v's degree: $f(v) = 1/(deg(v) - 1)$ (Newman 2001a, b). The motivation is to approximate the strength of the bond between any two authors. The total weight between any two nodes is then defined as the sum of the weights of all shared neighbors. Of course, any other function could be chosen as well to assign the weights, e.g., an inverse-quadratic function.

A third approach tries to approximate a saturating effect: it can be assumed that two authors already know each other quite well after they have written some papers together. The next co-authored paper might not deepen the strength of their relationship much more. Li et al. thus suggest to use a hyperbolic tangent function (Li et al. 2005). Again, any other function $g(coocc(v, w))$ depending on $coocc(v, w)$ could be used to assign weights between v and w.

All three methods make use of some arbitrary choice: either of a threshold or a function to evaluate whether the number of co-occurrences is high enough to imply an edge between the nodes. The motivation for a new method was to evaluate whether the co-occurrence between any two nodes is significant. To understand this, it is easiest to consider the co-occurrence of two vertices as a special case of an association rule in which the sets X and Y just consist of a single node. Then we can in principle use interestingness measures as the ones above to evaluate their co-occurrence. However, from the perspective of a classic network analytic viewpoint the co-occurrence of any two vertices in a bipartite network can also be considered as a special type of a *network motif* as described in the next section. As we will sketch in the following, the significance of network motifs can also easily be tested for.

3.3 Network motifs and their significance

Network motif analysis was introduced by Alon et al. and its application to biological data sets were reviewed in his book (Alon 2006; Milo et al. 2002). A network motif is a subgraph which occurs significantly more often in a given graph than in a suitable corresponding random graph. The suitability of a random graph is often debated and there are many different random graph models to choose from. A random graph is said to be *corresponding* to some given graph G if it maintains certain structural elements of G. In all cases, a corresponding random graph needs to maintain at least the same number of nodes and edges. Besides this, the edges are placed anew and uniformly at random. Next to the size of the graph, other structural elements might be maintained as well. Popular structural elements to maintain are for example the degree of each single vertex (Girvan and Newman 2002) or the number of cycles of length 3 or 4 (Milo et al. 2002).

It is convenient to define a set $\mathcal{G}(G)$ that contains all possible graphs that maintain the given structural elements of G and perturb all others. This set is called a *random graph model*. Most often, these sets are too large to enumerate all of the members from the random graph model. However, it is often possible to sample uniformly at random (u.a.r.) from the model. A well-known model is

[2] According to an interestingness measure called *conviction*.

the Erdős-Rényi graph model $\mathcal{G}(n,m)$ (Bollobás 2001), which consists of all graph which contain exactly m edges and n nodes. A graph can be sampled u.a.r. from $\mathcal{G}(n,m)$ by building a set of all possible pairs of nodes and drawing exactly m of these pairs u.a.r. Another important random graph model was introduced by Gilbert; it is called the $\mathcal{G}(n,p)$ model. For any given n and $0 \leq p \leq 1$, a sample from $\mathcal{G}(n,p)$ can be drawn u.a.r. by connecting any two nodes with probability p. For historical reasons, the latter model is often addressed as the Erdős-Rényi graph model.

By maintaining some structural elements and randomizing the rest, the occurrence of any network motif of interest can be compared to its occurrence in a sample of the chosen corresponding random graph model. If its occurrence is not significantly higher or lower than the one in the sample, it can be *explained* by the random graph model. For example, if a graph has 10 nodes and 30 edges, a clustering coefficient of 0.8 of a single vertex is not surprising but within the range of the possible fluctuations of a graph with this edge density. The edge density alone already explains the observation of a single vertex with a high clustering coefficient. If the occurrence of a network motif is approximately normally distributed, the *observed occurrence occ*(M) of the motif M can be evaluated by its z-score as described by Milo et al. (2002):

$$z\text{-score}(M) = \frac{occ(M) - \overline{occ_{\exp}(M)}}{\sigma_{\exp}(M)} \tag{9}$$

where $\overline{occ_{\exp}(M)}$ denotes the mean of the experimentally observed occurrence of M in the sample, and $\sigma_{\exp}(M)$ denotes its standard variation. A z-score of higher than 3.29 or lower than 3.29 has a probability $p < 0.001$ to occur and can thus be seen as a significant event.[3]

When network motif analysis was introduced (Milo et al. 2002) the paper was immediately followed by a discussion of the most suitable corresponding random graph model. As Artzy-Randrup et al. pointed out in their article, other random graph models than the most simple Erdős-Rényi graph model would result in very different significance levels of certain motifs (Artzy-Randrup et al. 2004). Thus, the choice of a suitable random graph model is crucial in the correct evaluation of the significance of a network motif.

In the following we will use the general approach of determining significant network motifs to build a one-mode projection of bipartite graphs and especially discuss the crucial role of choosing the best random graph model.

4 A systematic approach to one-mode projections

Our method to build a one-mode projection of bipartite graphs follows the systematic approach of evaluating the significance of the occurrence of network motifs. As sketched above, this is done by comparison with their occurrence in a suitable corresponding random graph models. The co-occurrence of any two nodes on one side of a bipartite graph is then just a special case of a network motif; interestingly, it is as well a special case of an association rule, in which X and Y only contain a single product (or the node which it represents). Based on this, we can compare the co-occurrence in a given network with its expected value in any suitable corresponding random graph model. The expected values in the chosen random graph model can then replace the respective terms in the various interestingness measures cited above. We thus propose the following systematic approach to one-mode projections:

4.1 Systematic approach to one-mode projections

1. Choose a connection pattern (or motif) M of v, w from the side of interest to nodes on the other side, e.g., the *co-occurrence coocc*(v, w).
2. Choose a suitable random graph model $\mathcal{G}$ for a given bipartite graph;
3. Compute the expected occurrence of pattern M in $\mathcal{G}$ or compute the mean of the observed occurrences of M in a sample of $\mathcal{G}$.
4. Quantify the interestingness of the occurrence of M in comparison with the expected occurrence of M by, e.g., dividing the former by the latter (*lift*), subtracting the latter from the former (*leverage*), or by computing the z-score.
5. Choose those pairs v, w with the highest interestingness and connect them in the one-mode projection.

For this article, we have chosen the *co-occurrence* as the motif of interest and the *leverage* as the measure of interestingness. The next step is then to choose a suitable random graph model to evaluate the significance of this motif. We will now show that the *simple statistical independence model* (SIM) on which the leverage and lift are based, corresponds to a random graph model which we will call the *simple bipartite random graph model* (SiBiRaG).

4.1.1 The simple independence model (SIM) and the simple bipartite random graph model (SiBiRaG)

As described above, the SIM assumes that the probabilities of buying single products are independent and that thus the event of buying two different products can be expressed by

[3] This is given under the assumption that the occurrence of M in the random graph model is normally distributed.

the product of the two respective probabilities in case of independence (Eqs. 7 and 8).

This model can be interpreted as the outcome of a simple bipartite random graph model in which the probability $P(p_i, u_x)$ that node p_i is connected to node u_x is given by $deg(p_i)/r$. I.e., this model assumes that every customer buys p_i with the same probability and that thus every node representing a customer has the same probability $deg(p_i)/r$ to be connected to some product node p_i. In this model, the probability $P(p_i, p_j)$ that two products p_i, p_j are bought by the same customer u_x is simply given by the product of the two probabilities $P(p_i, u_x)P(p_j, u_x)$. The expected number of co-occurrences $E\ [coocc_{\text{SIM}}(p_i, p_j)]$ in the simple bipartite random graph model (SiRaBiG) is then given by summing over all customers:

$$E[coocc_{\text{SiBiRaG}}(p_i,p_j)] = \sum_{u_x\in U} P(p_i,u_x)P(p_j,u_x) \tag{10}$$

$$E[coocc_{\text{SiBiRaG}}(p_i,p_j)] = \frac{deg(p_i)deg(p_j)}{r}. \tag{11}$$

It can now be seen that the expected support in this model, i.e., $E[coocc_{\text{SiBiRaG}}(p_i,p_j)]/r = E[supp_{\text{SiBiRaG}}(p_i,p_j)]$ is given by $\frac{deg(p_i)\cdot deg(p_j)}{r^2}$ just as in Eq. 7. From this, also Eq. 8 follows. We have now shown that SIM in the statistical analysis and SiBiRAG in the network motif analysis are absolutely equivalent. We can furthermore equate $E\ [coocc_{\text{SiBiRaG}}(p_i, p_j)]$ and $E\ [coocc_{\text{SIM}}(p_i, p_j)]$.

Observation 1 *The statistical independence model (SIM) underlying the leverage and lift can be identified with the simple bipartite random graph model (SiBiRaG) in the statistical analysis of bipartite network motifs.*

In SIM and SiBiRaG only three structural elements are maintained: the number of nodes n on both sides, the number of edges m between them, and the degree sequence L of the side of interest. Since the expected co-occurrence between any two nodes can be directly computed, it is not even necessary to sample from SiBiRaG. This is computationally very advantageous.

Regarding the identity between SIM and SiBiRaG, it is natural to introduce the z-score as a further interestingness measure of the co-occurrence:

$$z\text{-score}_{\text{SIM}} = \frac{coocc(x,y) - E[coocc_{\text{SIM}}]}{\sigma_{\text{SIM}}}, \tag{12}$$

where σ_{SIM} is the standard deviation of SIM.

Although SIM has the great advantage of yielding closed formulas for the expected values of the co-occurrence of two nodes, it is not the only possible random graph model that might be suitable. In general we want to use a random graph model that is still simple but maintains all structural elements that need to be maintained.

A very important structural element of many real-world complex networks is that degree distributions are skewed (Dorogovtsev and Mendes 2003). Figure 1 shows the degree distributions of 10 samples of 10,000 users each that rated a subset out of 17,770 films (description of the data set in Sect. 6.1). It can be clearly seen that the films' degree distribution and the users' degree distribution are very skewed in all data samples, i.e., most films are seen by a few users but some are seen by almost all. Vice versa, some users rate only a few films, but some of them rate almost all of them. If this is a persistent behavior of users, a natural random bipartite graph model is then the *fixed degree sequence model* which maintains the degree sequence of both sides of the graph as sketched in the following.

4.1.2 Fixed degree sequences model (FDSM)

Although SIM (and thus SiBiRaG) seems to be reasonable and is classically used, Gionis et al. argued that it is more appropriate to use a model conditional on both degree

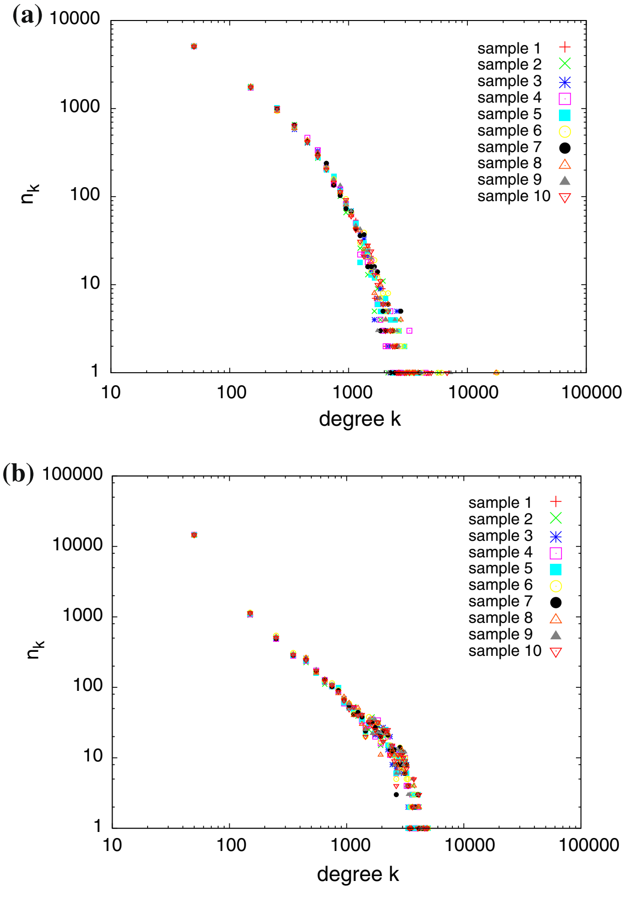


Fig. 1 **a** The degree distribution of users in 10 independent samples of 10,000 users that rated subsets of 17,770 films. **b** The according degree distributions of the films. n_k denotes the number of nodes with the given degree, binned into bins of size 100 each, and plotted at the mid of each bin

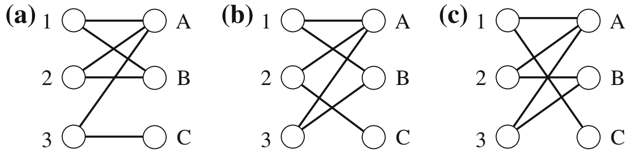


Fig. 2 $\mathcal{G}(L, R)$ for $L = \{2, 2, 2\}$ and $R = \{3, 2, 1\}$. It can be seen that *A* must be connected to all vertices on the *left-hand side*. *B* can be connected to any two vertices on the *left-hand side*, while *C* is connected to the left-over vertex. $E\ [coocc_{\mathrm{FDSM}}(i, j)]$ for any two vertices i, j on the left is 4/3. $E\ [coocc_{\mathrm{FDSM}}(A, B)] = 2$, $E\ [coocc_{\mathrm{FDSM}}(B, C)] = 0$, and $E\ [coocc_{\mathrm{FDSM}}(A, C)] = 1$

sequences (Gionis et al. 2007). We will denote this model by *FDSM*, the *fixed degree sequences model*. In this model, the *expected co-occurrence* $E\ [coocc_{\mathrm{FDSM}}(p_i, p_j)]$ is defined as

$$E[coocc_{\mathrm{FDSM}}(p_i, p_j)] = \frac{1}{|\mathcal{G}(L, R)|} \sum_{G \in \mathcal{G}(L,R)} coocc_G(p_i, p_j), \tag{13}$$

i.e., we sum over the real co-occurrences of p_i, p_j in all feasible graphs in $\mathcal{G}(L, R)$ and divide by the number of graphs in this set. A small example is given in Fig. 2.

Of course, statistical models should always reflect as much of a known structure as possible; they thus enable statements about new structures whose occurrence cannot be implied by all known structures. On the other hand, a too detailed model will reduce its applicability and it might also be computationally too expensive. The FDSM is clearly the better statistical model but it is also computationally quite expensive: without doubt it is impossible to enumerate all graphs in $\mathcal{G}(L, R)$ for even moderate size bipartite graphs.[4] Thus, we have to sample from $\mathcal{G}(L, R)$ with uniform probability and average the observed co-occurrences of all pairs of nodes in this sample to get an approximation of the expected co-occurrence. There are mainly two different approaches to sample from FDSM:

1. A Markov Chain Monte Carlo sampling as described in, e.g., (Brualdi 2006; Cobb and Chen 2003; Gionis et al. 2007; Holmes and Jones 1986).
2. Importance Sampling as described, e.g., in Chen et al. (2005) and Admiraal and Handcock (2008).

Note that the sampling itself is not the most time-consuming step: the main problem is to compute the co-occurrence values of all pairs of nodes for each sampled graph.

For the rest of the text we just assume that there is a method with which we can sample uniformly at random from FDSM and that we approximate expected values by averages of observed values in large enough samples; we thus apply the same approach as in the general network motif analysis. It is now obvious that the main advantage of SIM is that the expected values of co-occurrences can be directly computed, while the FDSM is a more detailed model but also computationally much more involved.

In the following we will argue why SIM is nonetheless not suitable for most real-world data sets and FDSM is the best alternative. Although Gionis et al. (2007) gave some anecdotal examples of why the FDSM is more appropriate, there is so far no theoretical result that shows why and when the SIM is not suitable. In the following, we extend their work by presenting theoretical observations of both models that show clearly why and when SIM fails.

5 Why SIM is not suitable for most real-world networks

Without loss of generality, we will be interested in the co-occurrence of products $p_i, p_j \in P$ in the *rows* of the 0-1 table, and thus on the nodes of the *left-hand side* of the respective bipartite graph, i.e., the set of products *P*. The results can be directly transferred to the co-occurrence of users by transposing the table.

5.1 Expected total number of co-occurrence events

We first determine the total number of co-occurrences of vertices in *P* for any given data set: each vertex u_x in *U* on the right-hand side of the graph induces $\binom{deg(u_x)}{2}$ co-occurrences because any two products in the user's market basket co-occur together. We denote the sum of all co-occurrences in graph *G* by *Coocc*(*G*):

$$Coocc(G) = \sum_{u_x \in U} \binom{deg(u_x)}{2}. \tag{14}$$

Since *Coocc*(*G*) is sufficiently defined by the degree sequence *R* alone, it can also be denoted by *Coocc*(*R*).

Coocc(*G*) = *Coocc*(*R*) can also be expressed in terms of *r*, *m*, and the variance *Var*(*R*) of the degree sequence. Note that

$$\sum_{u_x \in U} \binom{deg(u_x)}{2} = 1/2 \sum_{u_x \in U} deg(u_x)^2 - m/2$$

and

$$Var(R) = 1/r \sum_{u_x \in U} deg(u_x)^2 - \mu(R)$$

where $\mu(R)$ is the average degree in *R*. It follows that

[4] Note that the actual number $|\mathcal{G}(L, R)|$ of graphs in $\mathcal{G}(L, R)$ is not yet described by a closed formula (Greenhill and McKay 2008; Barvinok 2008).

$$Coocc(R) = \frac{r}{2} \cdot Var(R) + \frac{r}{2}\mu(R)^2 - m/2 \tag{15}$$

$$Coocc(R) = \frac{r}{2} \cdot Var(R) + \frac{m}{2}(\mu(R) - 1), \tag{16}$$

where the last equation follows from $\mu(R) = m/r$. It thus follows that even if m and r are fixed, the total co-occurrence $Coocc(R)$ is still linear in the variance of the degree sequence.

Let now L, R be given, and let $G = (L, R, E)$ be some graph of $\mathcal{G}(L, R)$.

Observation 2 *For any graph $G = (L, R, E)$, the sum*

$$\sum_{p_i \in L} \sum_{p_j \in L, j > i} coocc(p_i, p_j) = Coocc(R) \tag{17}$$

i.e., the sum of all co-occurrence values on the left -hand side has to equal the number $Coocc(G)$of co-occurrences induced by the right side.

Justification 1 *On both sides of the equation, we count the same quantity from two different perspectives: once from the perspective of the products and once from the perspective of the users. The total amount of co-occurrence events must of course be equal.*

It can be expected that any reasonable suitable random graph model needs to predict the correct total number of co-occurrence events. That is, we require that the sum of all *expected* co-occurrences of vertices in L for some given degree sequences L, R equals $Coocc(R)$. We will now show that this is the case for the FDSM:

Observation 3 *The sum of all* expected *co-occurrences of vertices in L in FDSM equals Coocc(R) for all degree sequences L, R.*

Proof

$$\sum_{p_i \in L} \sum_{p_j \in L, i \neq j} E[cooccFDSM(p_i, p_j)] \tag{18}$$

$$= \sum_{p_i \in L} \sum_{p_j \in L, i \neq j} \frac{1}{|\mathcal{G}(L, R)|} \sum_{G \in \mathcal{G}(L,R)} coocc(p_i, p_j) \tag{19}$$

$$= \frac{1}{|\mathcal{G}(L, R)|} \sum_{G \in \mathcal{G}(L,R)} \sum_{p_i \in L} \sum_{p_j \in L, i \neq j} coocc(p_i, p_j) \tag{20}$$

$$= \frac{1}{|\mathcal{G}(L, R)|} \sum_{G \in \mathcal{G}(L,R)} Coocc(R) \tag{21}$$

$$= Coocc(R) \tag{22}$$

It is, however, not true for *SiBiRaG* and thus for SIM:

Observation 4 *SIM implicitly claims that the degree sequence on the right-hand side can be approximated by a normal distribution.*

Justification 2 *The independence model assumes that each product has a probability of $deg(p_i)/r$ to be connected to any of the nodes representing the users. That means that each user has an expected degree of $\sum_{i \in L} deg(p_i)/r = m/r$, and the degree sequence can be approximated by a normal distribution for large data sets.*

In SIM all users have approximately the same degree and the total number of co-occurrence events only depends on the degree sequence of L:

Lemma 1 *The total number of co-occurrences of pairs p_i, p_j predicted by SIM is only depending on the degree sequence of L.*

Proof The sum of the expected co-occurrences E [$Coocc_{\mathrm{SIM}}(G)$] in SIM is given by

$$\frac{1}{2} \sum_{p_i \in L} \sum_{p_j \neq p_i \in L} E[coocc_{\mathrm{SIM}}(p_i, p_j)] \tag{23}$$

$$= \frac{1}{2} \sum_{p_i \in L} \sum_{p_j \neq p_i \in L} \frac{deg(p_i) deg(p_j)}{r} \tag{24}$$

$$= \frac{1}{2r} \sum_{p_i \in L} deg(p_i)(m - deg(p_i)) \tag{25}$$

$$= \frac{m^2}{2r} - \frac{1}{2r} \sum_{p_i \in L} deg(p_i)^2. \tag{26}$$

□

Equation 17 already showed that for any given graph G the real number of all co-occurrences of vertices in L equals $Coocc(R)$ and thus only depends on R. The higher the variance of R the larger is $Coocc(R)$. For large deviations between E [$Coocc_{\mathrm{SIM}}(G)$] and $Coocc(R)$, it is thus clear that the predictions by SIM must fail. This can be best shown on a worst-case example.

Theorem 1 *There is a family of bipartite graphs $G(n)$, representing n users and n products, such that the difference of the total number of co-occurrence events predicted by SIM and FDSM is in $\Omega(n^3)$.*

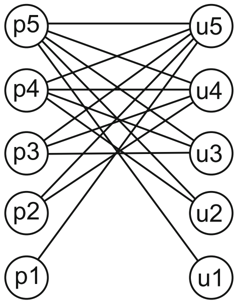


Fig. 3 An example of $G(n)$ with $n = 5$

Proof We first introduce the family $G(n)$, depicted in Fig. 3 for $n = 5$. In general, $G(n)$ is a bipartite graph between n products and n users, with $L = R = \{1, 2,\ldots, n\}$. A close inspection reveals that for every vertex there is exactly one subset of possible neighbors due to the constraint that no multi-edges are allowed. More generic, the only bipartite graph satisfying L, R is given by $E = \{\{u_x, p_i\} \mid 1 \le x \le i \le n\}$. It follows that the observed co-occurrence and the expected co-occurrence $E\,[cooc c_{\mathrm{FDSM}}(p_i, p_j)]$ is given by $\min\{i, j\}$.

$Coocc(R)$ is given by:

$$Coocc(R) = \sum_{i \in R} \frac{i(i-1)}{2} \tag{27}$$

$$= \frac{1}{2}\sum_{i \in R} i^2 - n(n+1)/2 \tag{28}$$

$$= \frac{n^3}{6} + O(n^2) \tag{29}$$

In SIM, the expected co-occurrence of two vertices p_i, p_j is given by ij/n. The total number of co-occurrence events $E\,[Coocc_{\mathrm{SIM}}(G)]$ in SIM is thus given by:

$$E[Coocc_{\mathrm{SIM}}(G)] = \sum_{i \in L} \sum_{j > i \in L} \frac{ij}{n} \tag{30}$$

$$= \frac{(n+1)}{2}\sum_{i \in L} i - \frac{1}{2n}\sum_{i \in L} i^3 - \frac{1}{2n}\sum_{i \in L} i^2 \tag{31}$$

$$= \frac{n^3}{8} + O(n^2). \tag{32}$$

Thus, the difference between the real number of co-occurrence events in $G(n)$ and the one predicted by SIM is in $\Omega(n^3)$. □

It now becomes obvious that the main problem of SIM is its underestimation of the total number of co-occurrence events when R shows a large variance. It follows that FDSM should be used whenever the predicted total number of co-occurrence events in SIM deviates strongly from the real number of co-occurrence events in the given data set:

Corollary 1 *Let $G = (L, R, E)$, where w.l.o.g. L contains the degrees of the side of interest. If the total number of co-occurrence events $Coocc(R)$ as described in Eq. 14 deviates from the expected total number $E\,[Coocc_{\mathrm{SIM}}(G)]$ of co-occurrence events in SIM as given by Eq. 23, SIM is not suitable as a null-hypothesis model. This is the case for most real-world data sets, since the deviation of the two values increases with the variance of R and most real-world networks show degree sequences with a large variance (Newman et al. 2006) (see Fig. 1).*

5.2 Contingency tables in SIM and FDSM

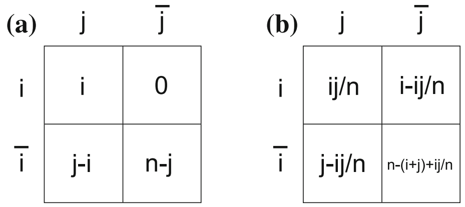
Another important statement concerns so-called *contingency tables*: for any given bipartite graph between products and consumers and any two products i and j, their contingency table presents how often both products are found in a basket (table entry ij), how often either of the products is in a basket (table entry $i\bar{j}$ and $\bar{i}j$), and how often none of them is contained in a basket (table entry $\overline{ij}$). Of course, all numbers add up to the number of baskets n. For the above given family $G(n)$ the expected values of all four cases can be computed and compared (see Fig. 4).

It can be seen immediately that the two models do not agree in any of the cases. Furthermore, their asymptotic behavior for $n \to \infty$ is also very different. In the following we will discuss the asymptotic behavior of leverage and lift.

5.3 Asymptotic behavior of leverage and lift in $G(n)$

It is also interesting to look at the asymptotic behavior of leverage and lift in $G(n)$. Since $L = R = 1, 2,\ldots, n$ allows for only one graph the co-occurrences between all pairs of nodes are just caused by the structure of the graph. Thus, the leverage of these co-occurrences should intuitively evaluate to 0 and lift should evaluate to 1, at least asymptotically. We add an index ∞ to the measures to indicate the behavior for the asymptotic limit of $n \to \infty$. We differentiate three cases regarding the degrees $deg(p_i) = i$ and $deg(p_j) = j$ of the chosen nodes p_i and p_j:

1. if i and j are constant, $i < j$,
   (a) $lev_{\mathrm{SIM}}, \infty(i,j) = i = O(1) \neq 0$ and
   (b) $lift_{\mathrm{SIM}}, \infty = n/j = O(n) \neq 1$.
2. For i, j proportional to n, i.e., $i = cn < j = c'n$, c, $c' \le 1$:
   (a) $lev_{\mathrm{SIM}}, \infty(i,j) = cn - cnc' = cn(1 - c') = O(n) \neq 0$ and



(a)

	j	$\bar{\text{j}}$
i	i	0
$\bar{\text{i}}$	j-i	n-j

(b)

	j	$\bar{\text{j}}$
i	ij/n	i-ij/n
$\bar{\text{i}}$	j-ij/n	n-(i+j)+ij/n

Fig. 4 General description of the expected number of baskets which contain both products i and j (ij), only product i but not j ($i\bar{j}$), only product j but not i ($\bar{i}j$), or neither of them ($\bar{i}\bar{j}$) (In the graph family $G(n)$ as described in Sect. 5.1 for $j > i$). **a** Shows the results for FDSM. **b** Shows the results for SIM

(b) $lift_{\text{SIM}}, \infty(i,j) = cn/(cnc') = 1/c' = O(1) \neq 1$.

3. When i is constant and $i < j = cn$, $c < 1$, then:

(a) $lev_{\text{SIM}}, \infty(i,j) = i - icn/n = i(1-c) = O(1) \neq 0$ and

(b) $lift_{\text{SIM}}, \infty(i,j) = 1/c = O(1) \neq 1$.

In summary, we have now stated that SIM can almost never be used for real data sets. However, sampling from FDSM is much more involved than computing expected co-occurrence values in SIM. Since the main problem of SIM seems to be its wrong estimation of $Coocc(R)$, an intuitive idea is that, for each pair of products, the two predictions might only differ by a scalar or by a scalar factor. This would make it possible to use SIM as an approximation for the more involved FDSM. Another intuition is that SIM could be at least used as a lower bound, since it seems to underestimate the expected co-occurrence in a random rewiring model in most cases. In the following we will relate the expected co-occurrences in SIM and FDSM theoretically and experimentally, and show that SIM cannot be used as any kind of approximation in most real-world data sets.

5.4 Analysis of the expected co-occurrence in FDSM and SIM

A simple observation regards known upper and lower bounds on the co-occurrence of any two vertices:

Observation 5 *For each pair of vertices i, j on the same side of a bipartite graph, their maximal co-occurrence is bounded by* $\min\{deg(i), deg(j)\}$ *and their minimal co-occurrence is bound by* $\max\{0, deg(i) + deg(j) - n\}$.

Justification 3 *Two articles cannot be bought more often together than any one of them was bought alone. Moreover, if both articles were bought by* $n + x$ *customers in sum, they must at least be bought together by x of them. Let now one node i be fixed with degree deg(i). Fig. 5 shows the lower and upper bounds of the co-occurrence of this vertex for some vertices with different degrees. Note that all co-occurrences in the graph family exemplified in* Fig. 3 *are correctly described by the upper bound.*

5.5 Using SIM as a lower bound on FDSM

At first glance it seems that the expected co-occurrence in FDSM always exceeds the expected co-occurrence in SIM. But although this is true for most real-world data sets, there are a few counterexamples: consider a graph where R contains only vertices with degree 1. Thus, the expected co-occurrence is 0 for all pairs of vertices on the left. In SIM, the expected co-occurrence is strictly above 0 for all pairs. In the extreme, if L contains only two vertices with degree k each, i.e., $m = r = 2k$, the expected co-occurrence of these two vertices is $k/2$ in SIM and 0 in FDSM.

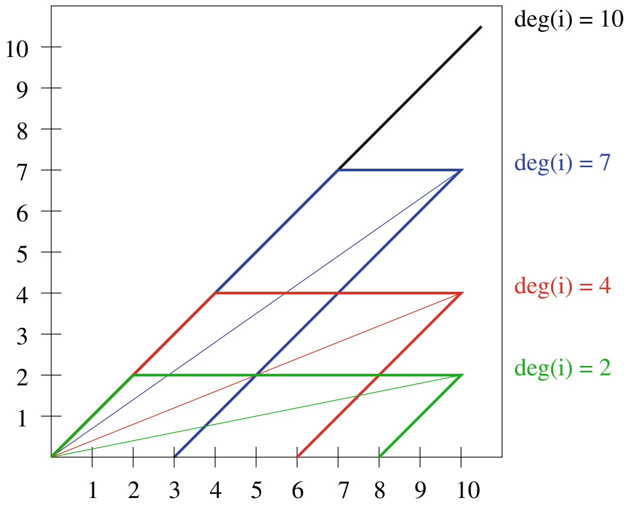


Fig. 5 Shown are the lower and upper bounds of co-occurrences for a vertex i with degree 2 (*green*), 4 (*red*), 7 (*blue*), or 10 (ten) in a bipartite graph in which there are 10 vertices on the other side. The x-axis denotes the degree of the other vertex j and the lines define the areas within which the co-occurrence of these two vertices must lie. The *diagonal* denotes $E\,[cooc_{\text{SIM}}(i, j)]$

In the examples seen so far, in a graph $G(L, R)$ with a smaller number of co-occurrence events $Coocc(R)$ than predicted by SIM, SIM gave an upper bound on all pairwise expected co-occurrences in FDSM. Thus, a straightforward question to ask is if $Coocc_{\text{SIM}} > Cooocc(R)$ ($Coocc\text{SIM} < Cooocc(R)$), are all single co-occurrences overestimated by the simple independence model? I.e., are the following implications correct?

$$Coocc_{\text{SIM}} > Cooocc(R) \rightarrow E[coocc_{\text{SIM}}(i,j)] > E[coocc_{\text{FDSM}}(i,j)]$$

$$Coocc_{\text{SIM}} < Cooocc(R) \rightarrow E[coocc_{\text{SIM}}(i,j)] < E[coocc_{\text{FDSM}}(i,j)]$$

for all $i, j \in L$

Although intuitive, this is not the case:

Lemma 2 *There is a family of graphs in which the co-occurrence of some pairs of vertices is underestimated by SIM while that of others is overestimated with respect to the predictions of FDSM.*

Proof Let G be a graph of three vertices on the left side with degrees $deg(1) = n - 1$, $deg(3) = n - 1$, and $deg(2) = n - 2$. Let R consist of $n - 2$ vertices with degree 3 and two vertices with degree 1 (see Fig. 6). It is obvious that the $n - 2$ vertices with degree 3 must be connected to all vertices in L, and that the two degree 1

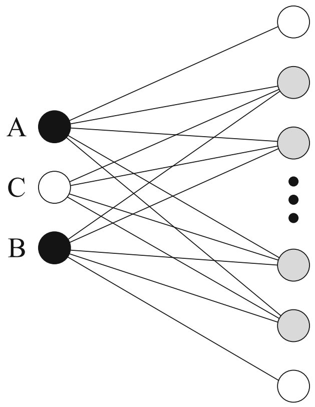


Fig. 6 The *grey* vertices on the left side have degree $n - 1$, the middle one has degree $n - 2$. On the *right-hand side*, there are $n - 2$ vertices with degree 3 (*striped*). These are thus connected to all vertices on the left. The remaining two *white* vertices are connected to one of the *grey* vertices each

vertices are then connected to one of the $n - 1$ degree vertices in L. Thus, there are two different feasible graphs in $\mathcal{G}(L, R)$.

SIM predicts a co-occurrence of $(n - 1)(n - 2)/n < n - 2$ for vertex pairs A, C and B, C and $(n - 1)(n - 1)/n > n - 2$ for vertex pair A, B. The expected co-occurrence in FDSM is $n - 2$ for all vertex pairs. Thus, some pairs' co-occurrence is overestimated and that of some underestimated in SIM with respect to FDSM. □

Note, however, that both models converge to an expected co-occurrence of $n - 2$ in the limit of large n.

5.6 Relationship between FDSM and SIM

The preceding observations and theoretical conclusions show that SIM is not suitable for degree sequences with a high variance. We have also shown that it cannot easily serve as a lower or upper bound on the expected co-occurrence in FDSM. We could not yet rule out that there is a simple approximative, linear relation between the two, i.e., that $E\ [cooc c_{\mathrm{FDSM}}(p_i, p_j)] \simeq k \cdot E\ [cooc c_{\mathrm{SIM}}(p_i, p_j)]$ for all p_i, p_j and some k. In the following we will show numerically that this is in general not the case.

To understand how $E\ [cooc c_{\mathrm{FDSM}}(i, j)]$ differs from $E\ [cooc c_{\mathrm{SIM}}(i, j)]$, we conducted experiments on a series of bipartite graphs with increasing variance in their degree sequences. Starting with a graph with 200 vertices on the left and 100 vertices on the right, we measured the co-occurrences of the vertices on the left. At the beginning, all vertices on the left had degree 10, and all vertices on the right had degree 20, m being 2000. Based on this, we skewed the degree sequences L, R in the following way: two vertices on the same side were chosen uniformly at random (u.a.r.) and the higher degree was incremented while the lower was decremented by one.[5] We performed this skewing operation for 0, $0.1m$, m, and $2m$ events. This resulted in four different pairs of L, R, ranging from totally uniform to strongly skewed sequences. For each of these pairs of L, R, we computed 50,000 sample graphs from $\mathcal{G}(L, R)$, each one computed by a random walk with $5m\log m = 76005$ steps from the preceding one. Then, we averaged over the observed co-occurrences for all pairs of vertices on the left. Figure 7 shows the difference between the average, observed co-occurrence $obs\ [cooc c(i, j)]$ and $E\ [cooc c_{SIM}(i, j)]$ for some fixed vertices i, i.e., it shows $obs\ [cooc c(i, j)] - E\ [cooc c_{\mathrm{SIM}}(i, j)]$ with fixed i vs j. Note that $obs\ [cooc c(i, j)]$ is used as an estimate for $E\ [cooc c_{\mathrm{FDSM}}(i, j)]$. Thus, if $E\ [cooc c_{\mathrm{FDSM}}(i, j)]$ and $E\ [cooc c_{\mathrm{SIM}}(i, j)]$ would only differ by a scalar or by a scalar factor, then $obs\ [cooc c(i, j)] - E\ [cooc c_{\mathrm{SIM}}(i, j)]$ should either be a constant or linear for fixed i and variable j.

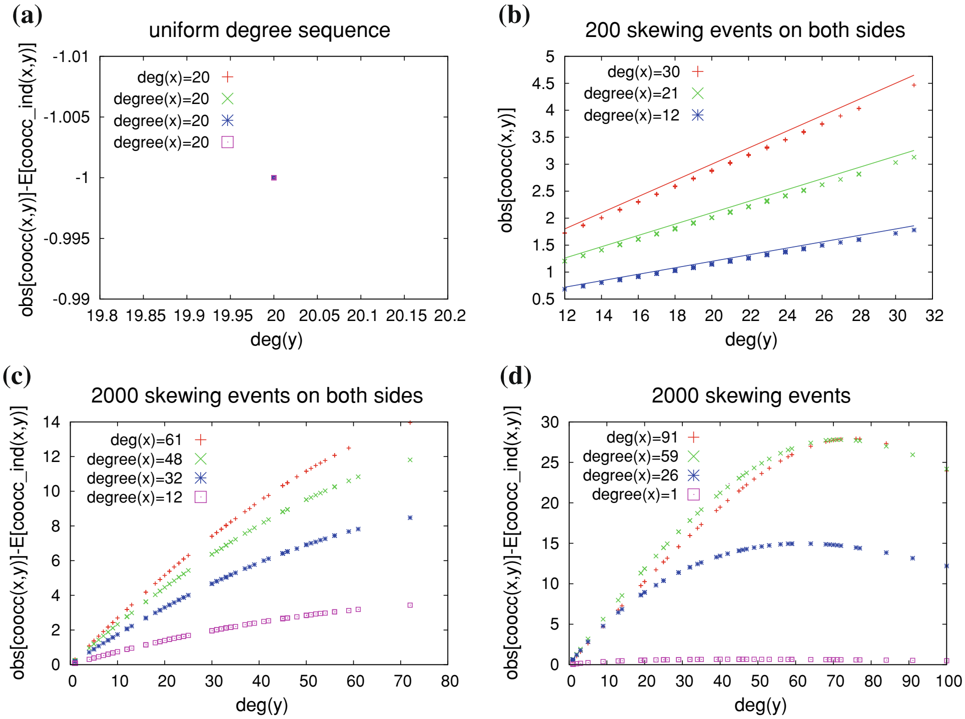
The first diagram shows that the observed co-occurrence for all pairs of vertices is essentially the same. It can be proven that this is always true, if the nodes on the side of interest all have the same degree:

Observation 6 *If all vertices in L have the same degree x, the expected co-occurrence of any two vertices $i, j \in L$ is* $E[cooc c_{\mathrm{FDSM}}(i,j)] = Cooc c(R)/\binom{l}{2}$.

Proof The first observation is that all pairs of vertices must have the same expected co-occurrence. This is because for each graph in which some pair i, j co-occurs more often than another pair i', j', there is an isomorphic graph G' in which i and i' and j and j' have switched their whole neighborhood. This defines a bijective mapping and thus for each graph in which $cooc c(i, j) - cooc c(i', j') = x$ there is another graph with $cooc c(i, j) - cooc c(i', j') = -x$. There are $\binom{l}{2}$ different pairs of vertices in $L \times L$. The sum of their expected co-occurrences must equal $Cooc c(R)$ by Lemma 3. It follows that $E[cooc c_{\mathrm{FDSM}}(i,j)] = Cooc c(R)/\binom{l}{2}$.

After the first $0.1m = 200$ skewing events, SIM overestimates the expected co-occurrence in FDSM for the exemplary vertices. And indeed, the real total number of co-occurrences is $Cooc c(R) = 9{,}475$, while SIM expects a total number of 19,797.6 co-occurrence events. After 2,000 skewing events, SIM starts to underestimate each single co-occurrence in FDSM. Most important to note is that the difference is largest for medium degrees, and the maximal difference seems to shift with increasing degree of x. It is easy to see that the expected co-occurrences in the two models are neither proportional to each other, nor do they have a constant difference. This implies directly that all

[5] As long as this does not lead to degree 0 or degree $r + 1$ $(l + 1)$ in which case nothing is done.



2000 skewing events

4000 skewing events

Fig. 7 Difference between the observed average co-occurrence sampled from $\mathcal{G}(L, R)$ and $E[coocc_{\mathrm{SIM}}(i, j)]$ for four randomly chosen vertices i vs. all other vertices j. **a** Shows results for a bipartite graph with uniform degree sequences, **b**, **c** and **d** show results for bipartite graphs with increasingly skewed degree sequences as described in the text

rankings of interestingness measures that are based on SIM need to be recomputed if the data set at hand is skewed enough. This has very important consequences, not only for the expected co-occurrence itself and thus also the support (Agrawal et al. 1993), *leverage* (Piatetsky-Shapiro 1991), or *lift* but also for other interestingness measures that implicitly rely on it, e.g., the *conviction* (Brin et al. 1997).

In essence, all of the above observations show that the co-occurrence of two events is simply not independent of the degree distribution on the right-hand side. The main point is that if A and B are bought together by some customer, and B and C as well, so are A and B. If furthermore there are customers that buy very many products, then there can be non-related items that are still bought together more often than expected in SIM. Our conclusion is that the first requirement of Piatetsky-Shapiro needs to be changed to

1. An interestingness measure $F(X,Y) = 0$ (or 1) if $P(X)$ and $P(Y)$ are conditionally independent given R and—if necessary—conditioned on other structural features of G as well.

This observation requires a generalized leverage and lift that replaces the *expected co-occurrence* term in SIM by that of FDSM or any other suitable corresponding random graph model. In the following, we will concentrate our discussions on FDSM. It is obvious that the new leverage $leverage_{\mathrm{FDSM}}$ and $lift_{\mathrm{FDSM}}$ fulfill the second requirement of Piatetsky and Shapiro, namely that the measures increase with increasing $P(XY)$ with fixed $P(X)$ and $P(Y)$. We will now show that also the third requirement is fulfilled.

5.7 Monotonicity of FDSM

The next theorem shows that the expected co-occurrence is monotonic in the FDSM, i.e., $E\ [coocc_{\mathrm{FDSM}}(X, Y)]$ increases monotonically with $P(X)$ (or $P(Y)$) when $P(XY)$ and $P(Y)$ (or $P(X)$) remain the same—thus $leverage_{\mathrm{FDSM}}$ and $lift_{\mathrm{FDSM}}$ decrease monotonically.

Theorem 2 *Monotonicity*

The expected co-occurrenceE $[coocc_{\mathrm{FDSM}}(x, y)]$ *is monotonic:E* $[coocc_{\mathrm{FDSM}}(x, y)] \geq E\ [coocc_{\mathrm{FDSM}}(v, w)]$ if $deg(x) \geq deg(v)$ and $deg(y) \geq deg(w)$.

Proof The proof proceeds as follows: we first claim that

$$E[coocc_{\mathrm{FDSM}}(x,z)] \geq E[coocc_{\mathrm{FDSM}}(y,z)] \ \mathit{if}\ deg(x) \geq deg(y).$$

Then, we conclude that if $deg(x) \geq deg(v)$ and $deg(y) \geq deg(z)$, then $E\ [coocc_{\mathrm{FDSM}}(x, y)] \geq E\ [coocc_{\mathrm{FDSM}}(v, y)] \geq E\ [coocc_{\mathrm{FDSM}}(v, w)]$ since the expected co-occurrence is a symmetric function.

To prove the claim, $\mathcal{G}(L, R)$ is partitioned into different subsets of graphs $G_{N',M'}(L, R)$. For each subset we show that the co-occurrence is strictly monotonic and thus the theorem is true for the whole set. Let now G be any graph in $\mathcal{G}(L, R)$, and denote by $N_G(x, \bar{y})$ the set of all neighbors of x not shared by *y* in *G*. Then, we define $N' = N_G(x, \overline{y}) \cup N_G(y, \overline{x})$, i.e., the symmetric difference of the sets of neighbors of *x* and *y*. Let M' denote the set of all edges minus the edges between *x* or *y* and a vertex from N'. The subset $\mathcal{G}_{N',M'}(L,R) \subseteq \mathcal{G}(L,R)$ is then defined as the set of all feasible graphs with fixed edge set M'. It is clear that this defines a valid partition since all graphs in $\mathcal{G}(L,R)$ are assigned to one subset and no graph is assigned to more than one subset. We are now interested in the sum of the co-occurrence of *x* and *y* with *z* within one given subset $\mathcal{G}_{N',M'}(L,R)$.

If *z* is not connected to any vertex in N', then *coocc*(*x*, *z*) = *coocc*(*y*, *z*) in all graphs in $\mathcal{G}_{N',M'}(L,R)$. Let now *G* be a graph with $coocc_G(x, z) < coocc_G(y, z)$ and observe the subset $\mathcal{G}_{N',M'}(L,R)$ induced by *G*. Since all edges from M' are fixed, we facilitate the argumentation by using a set of graphs G' consisting of only the vertices *x*, *y*, *z* and those from N', and the edges $E - M'$ between them. In the restricted graphs G', we will identify *x*, *y*, and *z* with their degrees w.r.t. N', i.e., $|N'| = x + y$. Note that each of the $\binom{x+y}{x} = \binom{x+y}{y}$ possible graphs is feasible, i.e., *x* can freely choose its neighbors in N' and *y* is then connected to the remaining vertices. We will now observe the sum of co-occurrences between *x*, *z* and *y*, *z* in $\mathcal{G}_{N'_G,M'_G}(L,R)$ from the perspective of *x*. Given *x*, *y* and *z*, and some feasible co-occurrence $q \leq \min\{y, z\}$ of *y* and *z*, the number of graphs *NG*(*coocc*(*y*, *z*) = *q*) in $\mathcal{G}_{N'_G,M'_G}(L,R)$ with this co-occurrence is exactly given by $\binom{z}{q}\binom{x+y-z}{y-q}$, and analogously the same is true for feasible co-occurrences between *x* and *z*.

Of course, if *y* co-occurs *q* times with *z*, then *x* co-occurs $z - q$ times with *z*. For $q \leq \lfloor z/2 \rfloor$, the claim thus follows trivially as the co-occurrence of *x* and *z* is at least as large as the co-occurrence of *y* and *z* in this case. Consider now the case $q > z/2$. We claim that *NG*(*coocc*(*x*, *z*)) ≥ *NG*(*coocc*(*y*, *z*)), i.e. $\binom{z}{q}\binom{x+y-z}{x-q} \geq \binom{z}{q}\binom{x+y-z}{y-q}$. Remember that $x \geq y$. Then the claim is true since

$$\binom{x+y-z}{x-q} > \binom{x+y-z}{y-q} \Leftrightarrow \frac{(x+y-z)!}{(x-q)!(y+q-z)!} > \frac{(x+y-z)!}{(y-q)!(x+q-z)!} \tag{33}$$

$$\Leftrightarrow \frac{(y-q)!(x+q-z)!}{(x-q)!(y+q-z)!} > 1 \tag{34}$$

$$\Leftrightarrow \frac{\Pi_{i=y+q-z+1}^{x+q-z} i}{\Pi_{j=y-q+1}^{x-q} j} > 1. \tag{35}$$

The last inequality is true since both products contain the same number $x - y - 1$ of factors, and since $q - z < z/2 < q$, the *i*th factor in the nominator is larger than the *i*th factor in the denominator, which completes the proof. □

With this last piece of information we can now propose new interestingness measures based on FDSM.

5.8 Generalizing leverage and lift

Here, we propose to generalize leverage and lift to evaluate the difference or ratio of the observed and the expected co-occurrence under any suitable corresponding random graph model. As we discuss later (see Sect. 7) there are even more involved random graph models than FDSM to consider. Of course, for any of them the monotonicity criterion would still have to be shown. For the rest of this article we will use the FDSM and denote the new leverage by $leverage_{\text{FDSM}}$ and lift by $lift_{\text{FDSM}}$:

$$leverage_{\text{FDSM}} = coocc(x,y) - E[coocc_{\text{FDSM}}(x,y)] \tag{36}$$

$$lift_{\text{FDSM}} = coocc(x,y)/E[coocc_{\text{FDSM}}(x,y)]. \tag{37}$$

Note that in contrast to SIM and $E\ [coocc_{\text{SIM}}(x, y)]$, there is no known closed formula for $E\ [coocc_{\text{FDSM}}(x, y)]$. Furthermore, the set $\mathcal{G}(L,R)$ will in most cases be by far too large to enumerate. Thus, $E\ [coocc_{\text{FDSM}}(x, y)]$ cannot be computed but needs to be estimated by computing the average co-occurrence of all pairs of vertices in a large enough sample from $\mathcal{G}(L,R)$. We will now discuss algorithmic aspects and experimental results using the new interestingness measure $leverage_{\text{FDSM}}$.

6 Experimental results

The proposed modified leverage $leverage_{\text{FDSM}}(v, w)$ basically defines a similarity measure, i.e., it assigns a real number to all pairs of nodes such that the more significant neighbors they have, the higher the number. We can now build two lists: A global list L_G containing triples of two products and their leverage can be built, sorted by the latter, or local lists *L*(*v*) for each node, containing all other products and their leverage to it, sorted by the latter. As with other similarity data, these lists can be used to turn the data into a sparse graph in many ways. E.g., the graph can be created from the best *O*(*n*) triples from the global list L_G or by connecting each node *v* to the nodes from the *k* highest-valued pairs in its local list *L*(*v*). Note that these edges are directed in the sense that *v* might find that *w* belongs to its *k* closest neighbors but not vice versa. By definition, both methods produce a sparse one-mode projection from any given data set. But of course the quality of this projection depends heavily on the quality of the similarity measure, the modified leverage $leverage_{FDSM}$. In the following we will describe how the quality of such a similarity measure can be assessed in general, and apply the methods to the modified leverage measure in particular.

6.1 Quality assessment

The main idea of our one-mode projection procedure is to create a sparse graph from a bipartite graph that can then be clustered by any reasonable clustering algorithm. In order to make the projection useful for clustering, most nodes (representing one object from the data set) should only be connected to nodes which represent similar objects. The proposed modified leverage assumes that the co-occurrence of two objects not only tells us about which products will be co-bought together or which authors are most likely to produce another article together, but also whether the products are similar by some intuitive measure and whether the authors of an article share some scientific interest. We conjecture that the modified leverage not only reliably picks the objects that co-occur significantly often in a stable and reliable way, but also that if we connect each object to the objects with which it most significantly co-occurs, these objects are also similar by content.

A general problem in the judgment of such a conjecture is that we seldom have something like a *ground truth* with which we can compare our results. Luckily, the Netflix prize data set (http://www.netflixprize.com) provides at least some possibilities to check the validity of the method. The data set consists of 100 million customer ratings of 17,770 films, between the 31st of December, 1999, and the 31st of December, 2005. There are over 480,000 distinct customers, identified by an ID between 1 and 2,600,000. The degree sequences of customers and films are both highly skewed as already seen in Fig. 1. For each rating event, the customer ID, the film ID, and the rating from 1 to 5 ('very bad' to 'very good') are presented. Additionally, a second file assigns to each film ID the film's title and its publishing year. The data allow for different quality assessment techniques:

1. Since the data set is so large, it can be partitioned randomly into smaller data sets and if the method is stable, all of the data sets should give rise to very similar rankings in L_G and $L(v)$.
2. We expect that rankings of high-degree nodes should be even more stable than those of low-degree nodes.
3. For films that are part of a series we expect that their best ranked neighbors are other parts in the same series and that all parts of the series are among the best ranked neighbors.

To analyze the stability of the given rankings, we used smaller samples from the data set. We computed 20 data samples DS_1 to DS_{20}, composed of all ratings of 10,000 users, each. The first data set contains the ratings of the 10,000 users with lowest IDs, the second all ratings of the next 10,000 users, and so on. For each sample, we computed the 1,000 pairs of films with globally highest leverage in a list, denoted by $L_G(DS_i)$. Since the leverage favors films with a high degree, we also computed for every data sample DS_i and every film v a local list $L(DS_i, v)$ containing its up to 100 best neighbors w sorted by their modified leverage $leverage_{FDSM}(v, w)$.

In the following, we will discuss some algorithmic aspects.

6.2 Algorithmic aspects

The procedure to compute $leverage_{\mathrm{SIM}}$ for all pairs of nodes of a given side in a bipartite graph is very simple: Given $G = (U \cup P, E)$ and being interested in the $leverage_{\mathrm{SIM}}$ values between the vertices in P, it is first necessary to compute the *observed* co-occurrence. For this, we initiate an array $\mathcal{C}occ\, of\, size |P| \times |P|$ and iterate over all nodes u_i in U. For each pair of neighbors p_x, p_y of u_i, we increase the according value $\mathcal{C}occ[x][y]$ by one. After iterating over all vertices $u_i, \mathcal{C}occ[x][y]$ contains $coocc(p_x, p_y)$, as shown in Sect. 5.1. The runtime is obviously in $O(Coocc(G) + |P|^2)$. For any sample with 20,000 users the empirical runtime of computing the co-occurrences of all pairs of vertices in P is 4.2 s ± 0.2 on an AMD Athlon$^{\mathrm{TM}}$ X2 Dual Core QL-65 with 2.1 GHz (only one processor was used). As a side effect, the degree of each vertex in P can be computed along with the former computation, and with this, $leverage_{\mathrm{SIM}}$ can afterwards be computed in $O(n^2)$.

Since there does not seem to be any closed formula for $E\, [cooocc_{\mathrm{FDSM}}(p_x, p_y)]$, it is necessary to sample from the according $\mathcal{G}(L, R)$, where L and R are determined by the degree sequences of vertices in U and P. We have implemented the simple Markov chain algorithm described by Gionis et al. (2007), in which at every time step a pair of edges $e_1 = (v,w)$ and $e_2 = (x,y)$ is drawn uniformly at random. If neither (v, y) nor (x, w) is in E, e_1 and e_2 are removed from E and (v, y) and (x, w) are added to E; this is called an *edge swap*. This process is repeated long enough to allow for *mixing*, that is, long enough such that the resulting graph G' is independent from the starting point G; in our experiments, we set the number of swapping steps to 70,000. The result of the first random walk was used as a starting point for the next random walk, and so forth. The theoretical runtime is directly proportional to the number of attempted swaps; the empirical runtime on the same machine as described above was 820 ms ± 40. For each of the 5,000 graphs that we sampled, we computed the co-occurrence for all pairs of vertices in U, with again an empirical runtime of 4.2 s ± 0.18. It thus becomes clear that the *sampling* from $\mathcal{G}(L, R)$ itself is not the bottleneck, but rather the computation of the co-occurrences in each of the samples. In sum, computing one sample and the co-occurrence of all pairs of vertices in it took around 5 s, and for all 5,000 samples it took roughly 7 h per data set.

If the leverage of any two films v, w is negative, this implies that they co-occur less often than expected, so we disregard neighbors with a negative leverage. Both, $L_G(DS_i)$ and $L(DS_i, v)$ give rankings. To compare rankings between different data samples, one can compute the percentage of objects that are listed in both rankings. To, moreover quantify the *order* in which the commonly listed objects are given, a *rank correlation coefficient* like Kendall's τ is needed, which will be described in the following.

6.3 Validating the ranking of a given similarity measure

To assess the stability of rankings given by some similarity measure like the leverage, Kendall's τ is a useful *rank correlation coefficient* (Kendall 1938). An easier formulation was given by Abdi (2007) on which we rely here. In its basic form it quantifies the correlation between two rankings on the same set of n objects, denoted by numbers of 1 to n. Given one ranking, which is w.l.o.g. represented by the sequence 1, 2, 3,…, n, the second ranking is then a simple permutation of these numbers. We will denote this second ranking by a function $\Pi(y)$ that gives the value on the y-th place in the second ranking. Vice versa, $\pi(x)$ denotes which place the x-th element of the first ranking has in the second ranking. E.g., let alon (2006), Admiraal and Handcock (2008), Agrawal et al. (1993), Abdi (2007); http://www.ninasnet.de/projects/omp_recommendations/10bestrecommendations.html) be the second ranking, then $\Pi(1) = 5, \Pi(2) = 3, \Pi(3) = 4, \Pi(4) = 2, \Pi(5) = 1$ and $\pi(1) = 5$, $\pi(2) = 4$, $\pi(3) = 2$, $\pi(4) = 3$, and $\pi(5) = 1$. To quantify the correlation between the two rankings, all ordered pairs of numbers in the second ranking are observed, i.e., (5, 3), (5, 4), (5, 2), (5, 1), (3, 4), (3, 2), (3, 1), (4, 2), (4, 1), and (2,1). A higher number followed by a smaller means that the respective objects had a different order in the first ranking. A pair (x, y) with $x > y$ is called a *discordant pair*, and the number of discordant pair of a ranking Π is denoted by $D(\Pi)$. Kendall's' τ is then defined as $1 - (4 \cdot D(\Pi)/(n(n-1))$ where n is the length of the ranking. It takes on values in $[-1, 1]$, where the extremes result for a reversed ranking ($\tau = -1$) and the same ranking ($\tau = 1$). For the above given example, Kendall's τ is thus $1 - 18/10 = -0.8$. Note that a slight change in the definition to $\sigma = 1 - 2 \cdot D(\Pi)/(n(n-1))$ equals the probability that any two pair of objects drawn u.a.r. have the same ordering in both rankings.

The main problem in computing Kendall's τ and its close cousin σ is to determine $D(\Pi)$. A naive implementation to compute $D(\Pi)$ has a runtime of $O(n^2)$ by checking every single pair. An improved algorithm with runtime $O(n \log n)$ was given by Newson (2006). However, in this special setting we expect the number of discordant pairs to be rather low. We will show that in this case, there is a more efficient algorithm to compute Kendall's τ that has a runtime of $O(n + D(\Pi))$, i.e., it is linear in the size of the ranking and bounded by above from the number of discordant pairs in the given permutation Π.

The algorithm walks through the second ranking Π and keeps two lists: After processing the i-th rank, *Bigger* contains all values $\Pi(i) > i$, i.e., those values in the rank that came earlier than in the first ranking, and *Smaller* contains all values i with $\pi(i) < i$, i.e., those values that are still missing. The values in *Bigger* have the same order as in Π and the values in *Smaller* are sorted in increasing order. With the help of these two lists, we count the number of discordant pairs. In essence, all elements in *Bigger* make for one discordant pair with each of the elements in *Smaller*. The algorithm guarantees that after the i-th rank is processed, all discordant pairs with (x, y) are accounted for, where $y \leq i$ and $\Pi(x) < i$.

Three cases have to be differentiated: if the current value in $\Pi(i)$ is equal to i, there is one discordant pair for each value $x \in$ *Bigger* and i, and one for i and all elements in *Smaller*. If $\Pi(i) \neq i$, we have to differentiate whether i has already been seen. If not, we add i to *Smaller*, because it is still missing, and thus, it produces a discordant pair with all elements in *Bigger*. If now $\Pi(i) > i$, all pairs $(\Pi(i), y)$ with $y \in$ *Smaller* are discordant. The element itself must be added to *Bigger* and it must be marked that the element has been seen before. If $\Pi(i) < i, \Pi(i)$ must be an element of *Smaller*. Let *Smaller* = [1, 4, 5, 7], and 4 the element at place i. Since 4 is placed at i, 4 will make a discordant pair with all elements in *Smaller* that are smaller than itself. Since *Smaller* is sorted increasingly, we can just walk through this list until we meet the respective element (in this case 4), remove it from *Smaller* because it is not missing anymore, and add one discordant pair for each element traversed so far. The last step that has to be done if $\Pi \neq i$, is to check whether i has been seen before. In that case, i is element of *Bigger* and needs to be removed. Since all elements in *Bigger* besides the element itself are at that time point larger than i and since the elements of *Bigger* have the same order as in $\Pi, \Pi(i)$ makes for a discordant pair with all elements in *Bigger* up until its own position in the list.

The runtime is in $\Omega(n)$ because each position is evaluated once. Whenever an element needs to be removed from *Bigger* or *Smaller*, the whole list might have to be traversed. Since every traversal in these lists stands for one discordant pair, the total runtime is bounded by $O(n + D(\Pi))$. In the worst case, i.e., a reverted ranking $\Pi = [n, \ldots, 3, 2, 1]$, the runtime is in $O(n^2)$. With this rank correlation coefficient, the different global and local

Table 1 Pairs of films with the 10 highest *leverage*$_{\mathrm{FDSM}}$ values in all 20 data samples

Lord of the Rings: The Two Towers	→	Lord of the Rings: The Fellowship of the Ring
Lord of the Rings: The Return of the King	→	Lord of the Rings: The Two Towers
Lord of the Rings: The Return of the King	→	Lord of the Rings: The Fellowship of the Ring
Lord of the Rings: The Fellowship of the Ring (Ext. Ed.)	→	Lord of the Rings: The Two Towers (Ext. Ed.)
Lord of the Rings: The Return of the King (Ext. Ed.)	→	Lord of the Rings: The Two Towers (Ext. Ed.)
Lord of the Rings: The Return of the King (Ext. Ed.)	→	Lord of the Rings: The Fellowship of the Ring (Ext. Ed.)
Star Wars: Episode VI: Return of the Jedi	→	Star Wars: Episode V: The Empire Strikes Back
Star Wars: Episode IV: A New Hope	→	Star Wars: Episode V: The Empire Strikes Back
Star Wars: Episode IV: A New Hope	→	Star Wars: Episode VI: Return of the Jedi
Kill Bill: Vol. 1	→	Kill Bill: Vol. 2

The *leverage*$_{FDSM}$ of all pairs is at least 725 in all data samples (© [2010] IEEE. From Zweig 2010, reprinted with permission)

rankings in the 20 data samples created from the Netflix data set can now be assessed.

In the following we will first describe experimental results on the stability of the specific global and local rankings in the 20 data samples. After that we will define a subset of films where the best recommendations are known and quantify how well the method agrees with them.

6.4 Global rankings

For each of the 20 data samples DS_i we computed the 1,000 pairs of films v, w with highest leverage. Note that out of the possibly more than 157,000,000 distinct pairs of films, the best ranked 1,000 films are less than 0.0007%. Already a very simple quality measure, which counts the number of common pairs of films for all data sets, reveals that the global rankings show a high overlap: restricted to the 10 highest-ranked pairs, all 20 (!) data samples list exactly the same pairs of films. These 10 pairs of films are displayed in Table 1.[6] Interestingly, these pairs give rise to three distinct 3-cliques and one single pair of films, and all films in the same component are obviously highly related: All Lord-of-the-Rings sequels are connected to each other, once in the normal cut and once in the extended version—but there is not yet a connection between the two components. Similarly, Star Wars Episode *IV* to *VI* are all pairwise related and thus build a triangle. The last pair makes a connection between both volumes of *Kill Bill*.

If more than the first 10 rankings are considered, the percentage of common pairs drops, as depicted in Fig. 8. But interestingly, the percentage of pairwise common pairs of films listed under the first k rankings seems to stabilize around 85%. As sketched above, this percentage is considerable compared with the enormous number of possible pairs of films in the data set. If we compute for each k the pairs of films that are listed under the k highest-ranked pairs in *all* 20 data samples (*maximal consensus*), this percentage seems to stabilize around 57.5%. I.e., given any $k \leq 1{,}000$, all data samples agree on around 57% pairs of films.

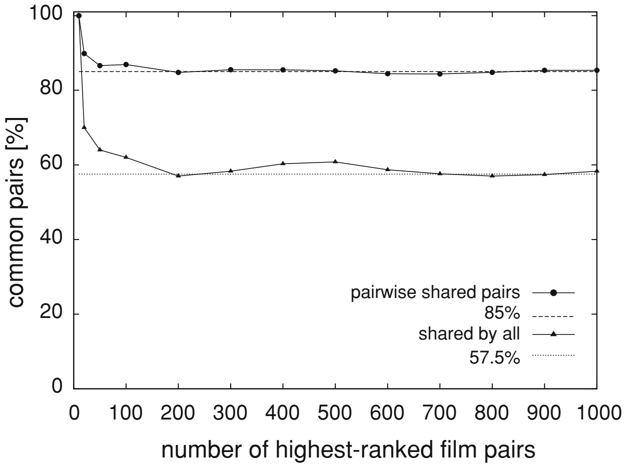


Fig. 8 Average percentage of pairwise shared pairs of films in $GL(DS_i)$ and $GL(DS_j)$, restricted to the first k rankings in all 20 data samples, and percentage of pairs of films listed under the first k rankings in all 20 data samples (maximal consensus). (© [2010] IEEE. Figure from Zweig 2010, reprinted with permission)

In the following, we restrict the rankings in each data set to the consensus pairs for a given k and compute Kendall's τ and σ. W.l.o.g., we set the ranking of the consensus pairs in the first data sample as reference and compare the rankings of all other data samples against it. Figure 9a shows that for the 10 highest ranked pairs (on which all data samples agree), Kendall's τ is on average 0.90, i.e., on average there are only 2 or 3 discordant pairs. Again, for higher k τ drops but seems to stabilize around 0.69. Note that the expected Kendall's τ is approximately normally distributed around 0 with a variance of $\sigma_\tau^2 = 2(2N+5)/(9{*}N{*}(N-1))$, with a satisfactory approximation for $N > 10$ (Abdi 2007). τ's significance can thus be tested by computing $Z_\tau = \tau/\sigma_\tau$, which denotes how many standard

[6] Note that the order was chosen for displaying reasons—none of the data samples directly showed them in this order.

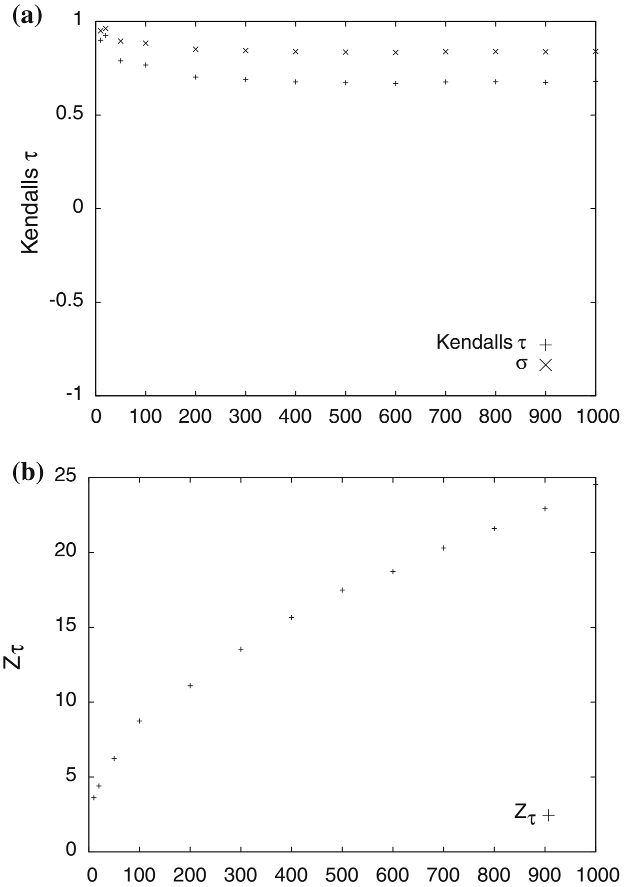


Fig. 9 Assessment of global rank correlation. **a** Average rank correlation (Kendall's τ) between first data sample and all other data samples with respect to the k first rankings. **b** Z_τ as defined in the text (© [2010] IEEE. Figure from Zweig 2010, reprinted with permission)

deviations the given τ value is away from the mean. Computing Z_τ for the average τ-values reveals that Z_τ increases from 4.4 for $k = 20$ to 24.5 ($k = 1{,}000$) and is thus highly significant (see Fig. 9b).

6.5 Local ranking

The last section showed strong evidence that the globally best pairs can reliably be found in quite small data samples of 10,000 users each. But of course, we are also interested in whether the method picks reliable recommendations for each single film. In the following we will show that the method can detect those objects of the data set for which the statistics is too poor to give any kind of recommendation. We think that this is a major advantage of the method, since giving no recommendation might be better than giving some random recommendation. For all other objects, the local rankings are similarly stable and reliable as the global ranking.

To assess the question for the validity of local rankings, we computed for each film v in each of the 20 data sets up to 100 other films w with highest $leverage_{\mathrm{FDSM}}(v, w)$. We restrict them to those w with $leverage_{\mathrm{FDSM}}(v, w) > 10$, i.e., we require that at least 10 more customers than expected rented these two films together. Let $lev_{10}(v)$ denote the set of other films w with $leverage_{\mathrm{FDSM}}(v,w) > 10$ in the given data set. For data sample 1, only 6931 out of 17770 films v have at least one neighbor in $lev_{10}(v)$. Then, analogously to the global rankings, we computed for each film v the consensus set of recommendations for all 20 data sets. If the consensus set $|cons(v)| > 2$, we computed the average Kendall's τ of all other data samples with respect to the ranking of data sample 1. In summary, for each film v in data set 1 we know how many customers rated it, i.e., its degree $deg(v)$, the maximal $leverage_{\mathrm{FDSM}}$ $lev_{max}(v)$ it has with any other film w, the number of neighbors w with at least $leverage_{\mathrm{FDSM}}(v,w) > 10$ (considering only the 100 highest values, denoted by $|lev_{10}(v)|$), the number of neighbors ranked by all data sets $|cons(v)|$, the average of Kendall's τ for the first ranking of the consensus set against all other 19 rankings, and the significance Z_τ of this value.

The first and rather intuitive result is that there is a positive correlation between the degree $deg(v)$ of a film v and its number of significant neighbors $|lev_{10}(v)|$ (see Fig. 10a). But especially among the low degree films, there are some with absolutely the same degree but very different numbers of significant neighbors: The films "Never Die Alone" and "Aqua Teen Hunger Force: Season 2" have both been rated 118 times, but the first has only 6 significant neighbors of which none is in the consensus set for all 20 data samples. The latter has 90 significant neighbors of which 33 are in the consensus set. Moreover, the rank correlation of these 33 consenting neighbors is more than significant with an average value of Kendall's $\tau = 60.84$ and $Z = 5.06$, i.e., the order in which these consenting films are given is significantly the same. This indicates that the $leverage_{\mathrm{FDSM}}$ of two films, if it is significant, is a reliable measure that will identify the same significant neighbors in different data sets.

The diagrams show in general that a small $leverage_{\mathrm{FDSM}}$ value is correlated with a low number of consensus neighbors. Thus, a low maximal $leverage_{\mathrm{FDSM}}$ $lev_{max}(v)$ indicates that the data sample is not good enough to make any statistically valid recommendations for film v, because the given recommendations strongly depend on the given data sample. Figure 10c, d finally shows that for all films whose maximal $leverage_{\mathrm{FDSM}}$ is at least 100, the consensus set almost always has at least 10 members and that the average ranking correlation coefficient is highly significant for them. This last bit of evidence shows that the proposed method enables the network analyst to assess whether the data sample at hand is good enough to give recommendations for any single object and second to give statistically reliable recommendations for those objects that have significant neighbors. Thus, any reasonable method that uses

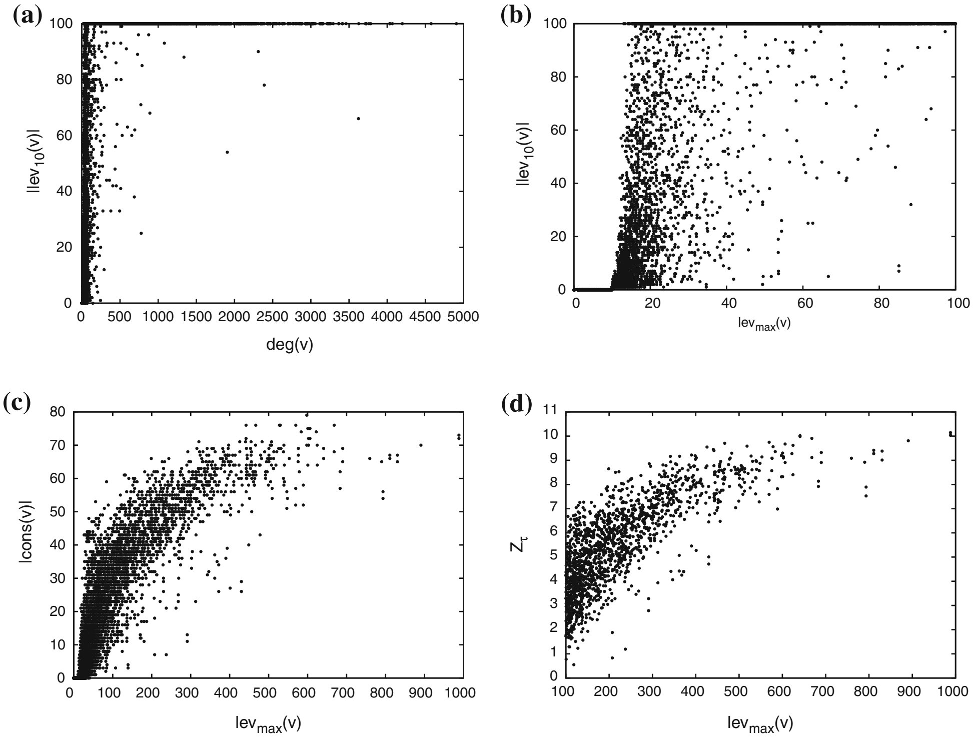


Fig. 10 Assessment of local ranking correlations. **a** Scatterplot of degree $deg(v)$ and number of significant neighbors $|lev_{10}(v)|$. **b** Scatterplot of $lev_{max}(v)$ and $|lev_{10}(v)|$ restricted to $lev_{max}(v) \in [0 : 100]$. **c** Scatterplot of $lev_{max}(v)$ and $|cons(v)|$. **d** Scatterplot of $lev_{max}(v)$ and Z_{tau} restricted to $lev_{max}(v) > 100$ (© [2010] IEEE. Figure from Zweig 2010, reprinted with permission)

the modified $leverage_{\mathrm{FDSM}}$ to build a one-mode projection of the bipartite graph will reliably connect those objects that are significantly co-occurring together. In the next section we will show for one subset of films that the method not only identifies objects that co-occur together significantly often but that these objects also have an objective similarity in the given Netflix data set.

6.6 Comparison of old and new leverage

As we have shown above, the old leverage definition ($leverage_{\mathrm{SIM}}$) suffered in most cases from extracting too many false positive rules. Table 2 shows all films in one of the datasets of 10,000 users which were rated by at least 3,000 users. The films show a wide range of genres, from melodrama to action, fantasy and animation, to documentaries, and romantic comedies. We grouped the films by hand, where, e.g., films 17–22 are typical romantic comedies, and films 36–43 are typical action films. It is rather intuitive that films within these groups should be co-rated significantly more often while films from one category are not expected to be significantly often co-rated with films from the other category. Broadly, we expect positive $leverage_{\mathrm{FDSM}}$ values within the group and negative values between them. However, the classic leverage $leverage_{\mathrm{SIM}}$ fails to indicate this (see Fig. 11). Since leverage values vary over a large scale, we plotted logarithmic values of the leverage. To deal with negative values, we plot $-\log(-\ leverage_X)$ with $X = FDSM$ or $X = SIM$. Negative values are encoded in red colors, positive values in blue. The classical leverage $leverage_{\mathrm{SIM}}$ indicates that all of the films are significantly often co-rated, against the intuition. The new leverage $leverage_{\mathrm{FDSM}}$ shows the full range of values: within the manually clustered groups the values are pairwise positive, while they are negative between most groups. There are some interesting results: ‘Sister Act’ goes quite well with the action film block, while ‘Pirates of the Caribbean’ rather behaves like one of the romantic comedies and not as a typical action film. And although the block from 6-16 shows films which are often co-rated in a quite coherent way, there are two exceptions: ‘American Beauty’ and ‘Shrek’, and ‘Shrek’ and ‘Pulp Fiction’. It is immediately obvious that ‘Shrek’ as a typical family movie might not share too much of an audience with the other two, while ‘American Beauty’ and ‘Pulp Fiction’ are liked by the same audience. This comparison already gives a good intuition about the new leverage. However, although the presented results are intuitive, it is hard to assess their absolute quality. In the following we will show results on a different subset, in which the quality of the given recommendations can be immediately assessed.

6.7 Benchmarking the quality

We have now shown that the method delivers very stable results, i.e., for all films v for which the data set contains

Table 2 Most popular films in a dataset comprising the ratings of 10,000 users, manually reordered and grouped. The numbers give the order in which the films are shown in the film-film-$leverage_{\mathrm{FDSM}}$-matrix in Fig. 11 (from left to right and bottom to top)

Number	Title	Netflix ID	Number of users that rated this film
1	The Royal Tenenbaums	8782	3,059
2	Lost in Translation	12232	3,152
3	National Treasure	17169	3,025
4	Troy	13081	3,020
5	I	5496	3,252
6	American Beauty	571	3,247
7	The Matrix	14691	3,017
8	Lord of the Rings: The Fellowship of the Ring	2452	3,145
9	Lord of the Rings: The Two Towers	11521	3,132
10	Shrek 2	3938	3,158
11	The Bourne Identity	6037	3,366
12	Bruce Almighty	3860	3,365
13	The Italian Job	4432	3,292
14	Pulp Fiction	11064	3,262
15	Ocean's Eleven	15107	3,414
16	Pirates of the Caribbean: The Curse of the Black Pearl	1905	4,068
17	50 First Dates	1962	3,012
18	Sister Act	6386	3,037
19	Two Weeks Notice	13050	3,115
20	How to Lose a Guy in 10 Days	14538	3,226
21	Sweet Home Alabama	15582	3,725
22	Dirty Dancing	7617	3,012
23	Gone in 60 Seconds	4996	3,125
24	Double Jeopardy	12911	3,149
25	The Day After Tomorrow	15205	4,036
26	Titanic	16879	3,042
27	Forrest Gump	11283	3,785
28	The Sixth Sense	4306	3,221
29	Gladiator	13728	3,132
30	Independence Day	15124	4,576
31	The Green Mile	16377	3,761
32	Top Gun	7624	3,211
33	What Women Want	2152	3,413
34	Miss Congeniality	5317	4,907
35	Pretty Woman	6287	4,034
36	Men of Honor	7234	3,160
37	Lethal Weapon 4	14367	3,234
38	Pearl Harbor	9340	3,624
39	The Rock	12317	3,461
40	Twister	12470	3,725
41	Con Air	16242	3,715
42	Armageddon	6972	3,628
43	The Patriot	14313	4,198

enough information, the method reliably assigns the same films w as most significant neighbors in all 20 data set samples. Moreover, it lists them in nearly the same order. But this does not yet imply that the most significant neighbors are also those films that are most similar with respect to the content. Of course, the latter is a necessary requirement to cluster the graph resulting from the one-mode projection. On the other hand, this aspect is in

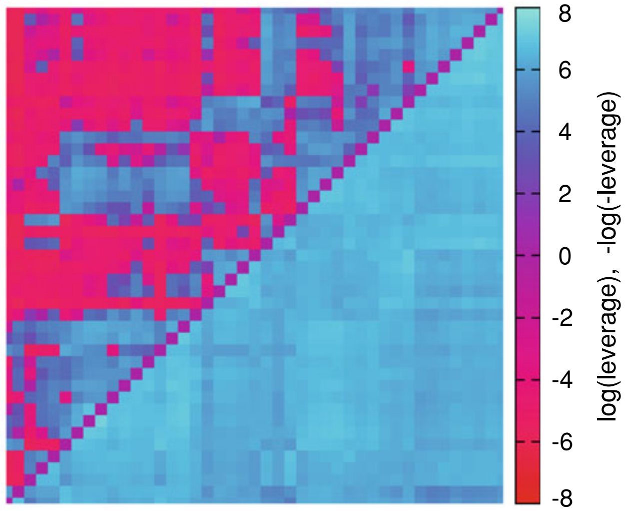


Fig. 11 Pairwise *leverage*$_{\mathrm{SIM}}$ (*lower-right triangle*) and *leverage*$_{\mathrm{FDSM}}$ (*upper-left triangle*) between all films that were rated at least 3,000 times in a set of 10,000 users. Plotted is $log(leverage_X)$ ($X = FDSM/SIM$) if $leverage_X > 0$, and $-log(-\,leverage_X)$ if $leverage_X < 0$. Films are assigned to columns from *left* to *right*, and to rows from *bottom to top* (see Table 2)

Table 3 Average quality assessment for recommendations of all sequels in a given series S sorted by *leverage*$_{\mathrm{SIM}}$ as described in the text with respect to the first data sample (first 20,000 subsequent user IDs)

Title of series S	n	pbr	pra	$\overline{first}$	$\overline{last}$
Northern Exposure	3	33.33	33.33	1.00	3.00
Seinfeld	3	33.33	0.00	–	–
Trailer Park Boys	3	0.00	0.00	–	–
Ren & Stimpy	3	0.00	0.00	–	–
Strangers with Candy	3	0.00	**100.00**	20.33	34.00
Survivor	3	0.00	0.00	–	–
The Dead Zone	3	0.00	33.33	33.00	72.00
The Jamie Kennedy Experiment	3	0.00	0.00	–	–
Roswell	3	33.33	33.33	5.00	23.00
Russell Simmons Presents Def Poetry	3	0.00	0.00	–	–
The Osbournes	3	0.00	33.33	2.00	12.00
The Shield	3	**100.00**	**100.00**	**1.00**	19.00
Sealab 2021	3	66.67	33.33	**1.00**	16.00
Silk Stalkings	3	33.33	33.33	**1.00**	6.00
SpongeBob SquarePants	3	0.00	0.00	–	–
Star Trek: Enterprise	3	0.00	33.33	17.00	64.00
24	3	**100.00**	**100.00**	**1.00**	15.33
Beast Wars Transformers	3	66.67	**100.00**	1.33	4.67
Boy Meets World	3	0.00	0.00	–	–
Cold Feet	3	33.33	33.33	**1.00**	**2.00**
ER	3	0.00	33.33	52.00	66.00
La Femme Nikita	3	33.33	33.33	14.00	15.00
Millennium	3	0.00	0.00	–	–
Monk	3	66.67	66.67	**1.00**	13.50
Yu-Gi-Oh!	3	0.00	33.33	2.00	11.00
In Living Color	4	0.00	25.00	14.00	35.00
Six Feet Under	4	75.00	50.00	**1.00**	12.00
Smallville	4	75.00	25.00	**1.00**	7.00
Profiler	4	50.00	25.00	**1.00**	20.00
Queer as Folk	4	**100.00**	75.00	**1.00**	35.67
Law & Order	4	25.00	50.00	21.50	79.50
Mr. Show	4	0.00	50.00	24.00	73.00
Alias	4	75.00	25.00	**1.00**	3.00
CSI	4	75.00	**100.00**	1.50	14.50
The West Wing	4	75.00	75.00	**1.00**	30.67
Will & Grace	4	0.00	25.00	8.00	95.00
Coupling	4	75.00	75.00	**1.00**	52.67
Curb Your Enthusiasm	4	50.00	25.00	**1.00**	45.00
Everybody Loves Raymond	4	0.00	0.00	–	–
The King of Queens	4	25.00	25.00	13.00	45.00
The Man Show	4	0.00	0.00	–	–
Farscape	4	50.00	75.00	3.33	23.33
Felicity	4	50.00	75.00	1.33	44.00
The Best of Friends	4	75.00	**100.00**	1.75	19.50
Gilmore Girls	4	75.00	25.00	**1.00**	12.00
King of the Hill	4	0.00	0.00	–	–
Andromeda	5	0.00	0.00	–	–
Oz	5	80.00	20.00	**1.00**	**4.00**
Angel	5	80.00	**100.00**	1.60	32.20
Babylon 5	5	80.00	**100.00**	2.00	19.40
Dawson's Creek	5	60.00	20.00	**1.00**	30.00
The Sopranos	5	**100.00**	40.00	**1.00**	34.00
Saved by the Bell: The New Class	5	20.00	0.00	–	–
A Touch of Frost	6	50.00	0.00	–	–
Dr. Quinn. Medicine Woman	6	0.00	16.67	26.00	60.00
Frasier	6	0.00	0.00	–	–
Hercules: The Legendary Journeys	6	0.00	0.00	–	–
Highlander	6	16.67	33.33	4.00	79.50
Homicide: Life on the Street	6	66.67	16.67	**1.00**	7.00
South Park	6	50.00	16.67	**1.00**	40.00
The Simpsons	6	66.67	16.67	**1.00**	47.00
Xena: Warrior Princess	6	0.00	33.33	5.50	73.00
Star Trek: Deep Space Nine	7	14.29	**100.00**	3.71	39.14
Star Trek: The Next Generation	7	28.57	71.43	3.20	27.60
Buffy the Vampire Slayer	7	85.71	**100.00**	3.43	23.43
Sex and the City	7	**100.00**	**100.00**	**1.00**	30.86
Star Trek: Voyager	7	85.71	85.71	**1.00**	19.00
Stargate SG-1	8	50.00	12.50	5.00	27.00
Friends	9	55.56	55.56	**1.00**	43.60
The X-Files	9	22.22	33.33	2.00	51.33

All optimal values are bold emphasised

Table 4 Average quality assessment for recommendations of all sequels in a given series *S* sorted by $leverage_{\mathrm{FDSM}}$ as described in the text with respect to the first data sample (first 20,000 subsequent user IDs)

Title of series *S*	*n*	*pbr*	*pra*	$\overline{first}$	$\overline{last}$
Northern Exposure	3	**100.00**	**100.00**	**1.00**	**2.00**
Seinfeld	3	**100.00**	**100.00**	**1.00**	**2.00**
Trailer Park Boys	3	0.00	0.00	–	–
Ren & Stimpy	3	0.00	33.33	5.00	24.00
Strangers with Candy	3	**100.00**	**100.00**	**1.00**	**2.00**
Survivor	3	33.33	**100.00**	1.67	8.67
The Dead Zone	3	66.67	**100.00**	1.33	3.00
The Jamie Kennedy Experiment	3	0.00	0.00	–	–
Roswell	3	**100.00**	**100.00**	**1.00**	5.67
Russell Simmons Presents Def Poetry	3	0.00	0.00	–	–
The Osbournes	3	**100.00**	**100.00**	**1.00**	2.33
The Shield	3	**100.00**	**100.00**	**1.00**	**2.00**
Sealab 2021	3	66.67	**100.00**	1.33	45.67
Silk Stalkings	3	66.67	66.67	**1.00**	9.00
SpongeBob SquarePants	3	33.33	33.33	16.00	27.00
Star Trek: Enterprise	3	66.67	**100.00**	2.33	29.00
24	3	**100.00**	**100.00**	**1.00**	**2.00**
Beast Wars Transformers	3	**100.00**	**100.00**	**1.00**	**2.00**
Boy Meets World	3	**100.00**	**100.00**	**1.00**	47.33
Cold Feet	3	**100.00**	**100.00**	**1.00**	3.67
ER	3	**100.00**	**100.00**	**1.00**	7.67
La Femme Nikita	3	**100.00**	**100.00**	**1.00**	4.33
Millennium	3	33.33	**100.00**	4.00	37.67
Monk	3	**100.00**	**100.00**	**1.00**	**2.00**
Yu-Gi-Oh!	3	33.33	**100.00**	13.00	36.00
In Living Color	4	**100.00**	25.00	**1.00**	4.00
Six Feet Under	4	**100.00**	**100.00**	**1.00**	3.75
Smallville	4	**100.00**	**100.00**	**1.00**	**3.00**
Profiler	4	**100.00**	**100.00**	**1.00**	13.75
Queer as Folk	4	**100.00**	**100.00**	**1.00**	**3.00**
Law & Order	4	75.00	**100.00**	1.50	5.50
Mr. Show	4	**100.00**	**100.00**	**1.00**	3.50
Alias	4	**100.00**	**100.00**	**1.00**	3.25
CSI	4	**100.00**	**100.00**	**1.00**	**3.00**
The West Wing	4	**100.00**	**100.00**	**1.00**	**3.00**
Will & Grace	4	**100.00**	**100.00**	**1.00**	16.25
Coupling	4	**100.00**	**100.00**	**1.00**	**3.00**
Curb Your Enthusiasm	4	**100.00**	**100.00**	**1.00**	3.50
Everybody Loves Raymond	4	**100.00**	50.00	**1.00**	22.50
The King of Queens	4	**100.00**	**100.00**	**1.00**	3.75
The Man Show	4	50.00	0.00	–	–
Farscape	4	**100.00**	**100.00**	**1.00**	**3.00**
Felicity	4	**100.00**	**100.00**	**1.00**	**3.00**
The Best of Friends	4	75.00	**100.00**	1.50	7.00
Gilmore Girls	4	**100.00**	**100.00**	**1.00**	3.25
King of the Hill	4	50.00	**100.00**	2.50	41.25
Andromeda	5	80.00	20.00	**1.00**	**4.00**
Oz	5	**100.00**	**100.00**	**1.00**	4.20
Angel	5	**100.00**	**100.00**	**1.00**	7.60
Babylon 5	5	**100.00**	**100.00**	**1.00**	4.40
Dawson's Creek	5	**100.00**	**100.00**	**1.00**	5.00
The Sopranos	5	**100.00**	**100.00**	**1.00**	**4.00**
Saved by the Bell: The New Class	5	60.00	40.00	**1.00**	10.00
A Touch of Frost	6	50.00	66.67	1.50	49.25
Dr. Quinn. Medicine Woman	6	66.67	16.67	2.00	27.00
Frasier	6	**100.00**	50.00	**1.00**	32.00
Hercules: The Legendary Journeys	6	50.00	0.00	–	–
Highlander	6	**100.00**	**100.00**	**1.00**	6.50
Homicide: Life on the Street	6	**100.00**	**100.00**	**1.00**	6.33
South Park	6	**100.00**	**100.00**	**1.00**	43.00
The Simpsons	6	**100.00**	**100.00**	**1.00**	13.33
Xena: Warrior Princess	6	**100.00**	**100.00**	**1.00**	5.33
Star Trek: Deep Space Nine	7	**100.00**	**100.00**	**1.00**	6.29
Star Trek: The Next Generation	7	**100.00**	**100.00**	**1.00**	**6.00**
Buffy the Vampire Slayer	7	**100.00**	**100.00**	**1.00**	**6.00**
Sex and the City	7	**100.00**	**100.00**	**1.00**	**6.00**
Star Trek: Voyager	7	**100.00**	**100.00**	**1.00**	**6.00**
Stargate SG-1	8	**100.00**	**100.00**	**1.00**	31.75
Friends	9	88.89	**100.00**	1.22	13.56
The X-Files	9	**100.00**	**100.00**	**1.00**	8.56

All optimal values are bold emphasised

general very hard to quantify objectively, as can be seen in the following examples which show four films and their two highest-ranked recommendations:

1. Dracula / The Strange Case of Dr. Jekyll and Mr. Hyde:
   (a) Dr. Jekyll and Mr. Hyde
   (b) Frankenstein / Bride of Frankenstein: The Legacy Collection
2. Frank Zappa: Does Humor Belong in Music?:
   (a) The Miles Davis Story
   (b) Frank Zappa: Baby Snakes
3. WWE: Summerslam 2004:
   (a) Wrestlemania XX 2004
   (b) WWE: Vengeance 2004
4. Gattaca:
   (a) The Fifth Element
   (b) Contact

Table 5 For each part in the series 'X-Files' we show the top five ranked other films according to *leverage*$_{\mathrm{FDSM}}$ (upper row) and *leverage*$_{\mathrm{SIM}}$ (lower row)

X-Files	1st	2nd	3rd	4th	5th
Season 1	The X-Files: Season 2	The X-Files: Season 3	The X-Files: Season 5	The X-Files: Season 6	The X-Files: Season 4
	Pirates of the Caribbean I	The Matrix	Lord of the Rings: The Fellowship of the Ring	Raiders of the Lost Ark	Independence Day
Season 2	The X-Files: Season 1	The X-Files: Season 3	The X-Files: Season 5	The X-Files: Season 4	The X-Files: Season 6
	The Matrix	Pirates of the Caribbean I	Lord of the Rings: The Fellowship of the Ring	Independence Day	Raiders of the Lost Ark
Season 3	The X-Files: Season 2	The X-Files: Season 1	The X-Files: Season 5	The X-Files: Season 4	The X-Files: Season 6
	The X-Files: Season 2	The Matrix	Pirates of the Caribbean I	The X-Files: Season 1	Raiders of the Lost Ark
Season 4	The X-Files: Season 2	The X-Files: Season 3	The X-Files: Season 5	The X-Files: Season 1	The X-Files: Season 6
	The Matrix	Pirates of the Caribbean I	Independence Day	The Silence of the Lambs	The Sixth Sense
Season 5	The X-Files: Season 2	The X-Files: Season 3	The X-Files: Season 1	The X-Files: Season 6	The X-Files: Season 4
	Pirates of the Caribbean I	The Matrix	Independence Day	Lord of the Rings: The Fellowship of the Ring	Raiders of the Lost Ark
Season 6	The X-Files: Season 2	The X-Files: Season 5	The X-Files: Season 1	The X-Files: Season 3	The X-Files: Season 7
	Pirates of the Caribbean I	Lord of the Rings: The Fellowship of the Ring	The Matrix	The X-Files: Season 2	Independence Day
Season 7	The X-Files: Season 6	The X-Files: Season 5	The X-Files: Season 2	The X-Files: Season 1	The X-Files: Season 8
	The X-Files: Season 6	The X-Files: Season 2	The X-Files: Season 5	Pirates of the Caribbean I	The X-Files: Season 1
Season 8	The X-Files: Season 5	The X-Files: Season 6	The X-Files: Season 2	The X-Files: Season 1	The X-Files: Season 3
	The Matrix	Pirates of the Caribbean I	The X-Files: Season 2	The X-Files: Season 5	The X-Files: Season 1
Season 9	The X-Files: Season 6	The X-Files: Season 8	The X-Files: Season 7	The X-Files: Season 2	The X-Files: Season 1
	Pirates of the Caribbean I	The X-Files: Season 2	The X-Files: Season 6	Indiana Jones and the Last Crusade	The Matrix

All of these recommendations seem to be reasonable and some are even interesting and non-obvious. But given the other 17,769 films in the data set it is hard to judge whether these are really the *best* recommendations. So, we wanted to find a set with something like a ground truth. Luckily, the data set at hand allows for some quality measure in this realm by concentrating on film series like Friends, or Star Trek. To find them, we have extracted all film titles that had the key word 'Season' in them, standing for one part of a series. We kept all series that had at least one volume that was published in 1990 or later (see Table 4 for an overview, data on series with less than 3 parts omitted). Given one film x out of a series and its list $L(x)$ of the first 100 highest-ranked films, we require that the highest-ranked film in $L(x)$ should be some part of the same series. Moreover, if the method is good, *all* other parts of the series should be recommended in the 100 most highly ranked films. To analyze the hypothesis, we computed for each of the series S

1. the number $n(S)$ of sequels in it;
2. the average percentage of films $pbr(S)$ for which the highest-ranked recommendation is another part from the same series; 100% is optimal;
3. the average percentage of films $pra(S)$ for which all parts from the same series were listed under the 100 highest-ranked recommendations (again, 100% is optimal);
4. among those films that list all other parts of the same series we computed

Table 6 For each part in the series ‘Friends’ we show the top five ranked other films according to *leverage*$_{\mathrm{FDSM}}$ (upper row) and *leverage*$_{\mathrm{SIM}}$ (lower row)

Series: Friends	1st	2nd	3rd	4th	5th
Season 1	Friends: Season 4	Friends: Season 3	The Best of Friends: Season 2	The Best of Friends: Season 1	Friends: Season 5
	Miss Congeniality	Forrest Gump	Pretty Woman	Pirates of the Caribbean I	Friends: Season 3
Season 2	Friends: Season 3	Friends: Season 1	Friends: Season 4	The Best of Friends: Vol. 1	Friends: Season 5
	Miss Congeniality	Forrest Gump	Pretty Woman	Pirates of the Caribbean I	Independence Day
Season 3	Friends: Season 1	Friends: Season 4	The Best of Friends: Season 2	Friends: Season 2	Friends: Season 5
	Miss Congeniality	Forrest Gump	Friends: Season 1	Pirates of the Caribbean I	Pretty Woman
Season 4	The Best of Friends: Season 2	The Best of Friends: Season 1	Friends: Season 1	Friends: Season 5	The Best of Friends: Season 3
	Miss Congeniality	Friends: Season 1	Forrest Gump	The Best of Friends: Season 2	Pretty Woman
Season 5	Friends: Season 6	The Best of Friends: Season 4	Friends: Season 4	Friends: Season 7	Friends: Season 8
	Friends: Season 4	Friends: Season 6	Miss Congeniality	Pretty Woman	The Best of Friends: Season 4
Season 6	Friends: Season 5	The Best of Friends: Season 4	Friends: Season 7	Friends: Season 8	Friends: Season 4
	Friends: Season 5	The Best of Friends: Season 4	Friends: Season 7	Friends: Season 4	Forrest Gump
Season 7	Friends: Season 6	Friends: Season 5	Friends: Season 8	The Best of Friends: Season 4	Friends: Season 4
	Friends: Season 6	Friends: Season 5	Miss Congeniality	Forrest Gump	Pretty Woman
Season 8	Friends: Season 6	Friends: Season 5	Friends: Season 7	The Best of Friends: Season 4	Friends: Season 4
	Friends: Season 5	Friends: Season 6	Friends: Season 7	Miss Congeniality	Forrest Gump
Season 9	Friends: Season 7	Friends: Season 6	Friends: Season 8	Friends: Season 5	The Best of Friends: Season 4
	Friends: Season 7	Friends: Season 6	Miss Congeniality	Pretty Woman	Forrest Gump

(a) the average rank $\overline{first}(S)$ of the first listed sequel from the same series (1 is optimal)

(b) the average rank $\overline{last}(S)$ of the last listed sequel from the same series ($n(S) - 1$ is optimal).

Again, we compare the results based on *leverage*$_{\mathrm{SIM}}$ and *leverage*$_{\mathrm{FDSM}}$: Table 3 shows the quality measures based on *leverage*$_{\mathrm{SIM}}$; all optimal results are shown in bold emphasised. It can be easily seen that there is not a single series in which all quality measures are optimal, but there are three in which 3 out of 4 measures are optimal. For more than half of the series, none of the values is optimal. Table 4 lists the same quality measures for the rankings based on *leverage*$_{\mathrm{FDSM}}$. Here, 19 of the listed 70 series are optimal with respect to all values, the recommendations of most series are optimal and near-optimal to 3 out of 4 values, and only 8 are non-optimal with respect to all values. In summary, the method has performed very well on this subset of assessable films which raises the hope that other, less-well assessable recommendations are of similar quality.

We have also computed the top ten ranked films for each film in a series. Due to space restricitions we only show the top five for all series with at least 7 parts in Tables 5–12; the full tables can be found at our website (http://www.ninasnet.de/projects/omp_recommendations/10bestrecommendations.html). It can be seen that the top five are more often from the same series using *leverage*$_{\mathrm{FDSM}}$ than if using *leverage*$_{\mathrm{SIM}}$. In general (looking at the whole set of series), it can be seen that *leverage*$_{\mathrm{SIM}}$ often ranks block busters very high—the ten films most often ranked under the first ten are Pirates of the Caribbean: The Curse of the Black Pearl; Independence Day; The

Table 7 For each part in the series ‘Stargate SG-1’ we show the top five ranked other films according to $leverage_{\mathrm{FDSM}}$ (upper row) and $leverage_{\mathrm{SIM}}$ (lower row)

Series: Stargate SG-1	1st	2nd	3rd	4th	5th
Season 1	Stargate SG-1: Season 2	Stargate SG-1: Season 3	Stargate SG-1: Season 4	Stargate SG-1: Season 6	Stargate SG-1: Season 5
	Independence Day	Armageddon	Pirates of the Caribbean I	Men in Black II	Lord of the Rings: The Fellowship of the Ring
Season 2	Stargate SG-1: Season 3	Stargate SG-1: Season 4	Stargate SG-1: Season 1	Stargate SG-1: Season 6	Stargate SG-1: Season 5
	Independence Day	Stargate SG-1: Season 3	Stargate SG-1: Season 1	Men in Black II	The Matrix
Season 3	Stargate SG-1: Season 2	Stargate SG-1: Season 4	Stargate SG-1: Season 6	Stargate SG-1: Season 5	Stargate SG-1: Season 1
	Independence Day	Stargate SG-1: Season 2	Stargate SG-1: Season 4	Men in Black II	Stargate SG-1: Season 6
Season 4	Stargate SG-1: Season 6	Stargate SG-1: Season 5	Stargate SG-1: Season 3	Stargate SG-1: Season 2	Stargate SG-1: Season 7
	Stargate SG-1: Season 6	Independence Day	Stargate SG-1: Season 5	Stargate SG-1: Season 3	Stargate SG-1: Season 2
Season 5	Stargate SG-1: Season 6	Stargate SG-1: Season 4	Stargate SG-1: Season 3	Stargate SG-1: Season 2	Stargate SG-1: Season 7
	Stargate SG-1: Season 6	Stargate SG-1: Season 4	Independence Day	Stargate SG-1: Season 3	Stargate SG-1: Season 2
Season 6	Stargate SG-1: Season 4	Stargate SG-1: Season 5	Stargate SG-1: Season 3	Stargate SG-1: Season 2	Stargate SG-1: Season 7
	Stargate SG-1: Season 4	Stargate SG-1: Season 5	Independence Day	Stargate SG-1: Season 3	Stargate SG-1: Season 2
Season 7	Stargate SG-1: Season 5	Stargate SG-1: Season 6	Stargate SG-1: Season 4	Stargate SG-1: Season 3	Stargate SG-1: Season 2
	Stargate SG-1: Season 5	Independence Day	Stargate SG-1: Season 6	Stargate SG-1: Season 4	Stargate SG-1: Season 3
Season 8	Stargate SG-1: Season 7	Stargate SG-1: Season 5	Stargate SG-1: Season 6	Stargate SG-1: Season 4	Stargate SG-1: Season 2
	Independence Day	Men in Black II	Lord of the Rings: The Two Towers	Lord of the Rings: The Fellowship of the Ring	Stargate SG-1: Season 7

Matrix; Lord of the Rings: The Fellowship of the Ring and *The Two Towers; Forrest Gump; Miss Congeniality; Spider Man; The Sixth Sense* and *American Beauty*. With $leverage_{\mathrm{FDSM}}$, the ten films most often ranked under the first ten are *Friends, Season 4,5,6,7; The Best of Friends, Season 4; CSI, Season 1 and 3; Buffy the Vampire Slayer 5 and 7.*. A *one-mode projection* of the 400 films that are part of a series and in which each film is connected to all films in its top ten ranks, results in a graph with 758 nodes for $leverage_{\mathrm{SIM}}$ and in a graph with 958 nodes for $leverage_{\mathrm{FDSM}}$. Although this might at first sound non-intuitive, $leverage_{\mathrm{FDSM}}$ identifies more different films as similar to the series than $leverage_{\mathrm{SIM}}$ which mainly points to block busters. A good example for this finding is the series “The World Poker Tour” with two seasons. $leverage_{\mathrm{FDSM}}$ connects them with films called “World Poker Tour: Battle of Champions”, “Winning Strategies: Texas Hold’em Poker with Mike Caro”, “Masters of Poker: Vol. 1: Phil Hellmuth’s Million Dollar Poker System”, and“Masters of Poker: Vol. 2: Phil Hellmuth’s Million Dollar Secrets to Bluffing and Tells”. $leverage_{\mathrm{SIM}}$ ranks none of these films among the first ten rankings, but rather films like “Jurassic Park” or “Enemy of the State”.

Figures 12 and 13 show two OMPs on a restricted subsets of films: in both cases, we started with a graph in which every part of a series was connected to its top ten ranked other films, as indicated by $leverage_{\mathrm{SIM}}$ and $leverage_{\mathrm{FDSM}}$, respectively. We kept only those edges that were reciprocal and show all connected components with at least four nodes, since otherwise the graphs contained too many components to be visualized on one page. It can be seen that the OMP based on $leverage_{\mathrm{SIM}}$ only contains 68

Table 8 For each part in the series 'Star Trek: Voyager' we show the top five ranked other films according to *leverage*$_{\mathrm{FDSM}}$ (upper row) and *leverage*$_{\mathrm{SIM}}$ (lower row)

Series: Star Trek: Voyager	1st	2nd	3rd	4th	5th
Season 1	Star Trek: Voyager: Season 2	Star Trek: Voyager: Season 3	Star Trek: Voyager: Season 4	Star Trek: Voyager: Season 5	Star Trek: Voyager: Season 6
	Independence Day	Pirates of the Caribbean I	The Matrix	Indiana Jones and the Last Crusade	Spider-Man
Season 2	Star Trek: Voyager: Season 1	Star Trek: Voyager: Season 3	Star Trek: Voyager: Season 4	Star Trek: Voyager: Season 5	Star Trek: Voyager: Season 6
	Star Trek: Voyager: Season 1	Independence Day	Pirates of the Caribbean I	Star Trek: Voyager: Season 3	Indiana Jones and the Last Crusade
Season 3	Star Trek: Voyager: Season 2	Star Trek: Voyager: Season 4	Star Trek: Voyager: Season 1	Star Trek: Voyager: Season 5	Star Trek: Voyager: Season 6
	Star Trek: Voyager: Season 2	Star Trek: Voyager: Season 4	Star Trek: Voyager: Season 1	Star Trek: Voyager: Season 5	Star Trek: Voyager: Season 6
Season 4	Star Trek: Voyager: Season 3	Star Trek: Voyager: Season 2	Star Trek: Voyager: Season 5	Star Trek: Voyager: Season 6	Star Trek: Voyager: Season 1
	Star Trek: Voyager: Season 3	Star Trek: Voyager: Season 2	Independence Day	Star Trek: Voyager: Season 5	Indiana Jones and the Last Crusade
Season 5	Star Trek: Voyager: Season 6	Star Trek: Voyager: Season 4	Star Trek: Voyager: Season 3	Star Trek: Voyager: Season 7	Star Trek: Voyager: Season 2
	Star Trek: Voyager: Season 6	The Matrix	Indiana Jones and the Last Crusade	Star Trek: Voyager: Season 4	Star Trek: Voyager: Season 3
Season 6	Star Trek: Voyager: Season 7	Star Trek: Voyager: Season 5	Star Trek: Voyager: Season 4	Star Trek: Voyager: Season 3	Star Trek: Voyager: Season 2
	Star Trek: Voyager: Season 7	Star Trek: Voyager: Season 5	Indiana Jones and the Last Crusade	Independence Day	Star Trek: Voyager: Season 4
Season 7	Star Trek: Voyager: Season 6	Star Trek: Voyager: Season 5	Star Trek: Voyager: Season 4	Star Trek: Voyager: Season 3	Star Trek: Voyager: Season 1
	Star Trek: Voyager: Season 6	Indiana Jones and the Last Crusade	The Matrix	Star Trek: Voyager: Season 5	Pirates of the Caribbean I

of the starting 400 films and parts of only 11 series of the original 165 series. The only non-trivial component shows a connection between the series "Friends" and "The best of Friends". The OMP based on *leverage*$_{\mathrm{FDSM}}$ contains 276 out of the 400 films we started with, and 49 out of the 165 series. Some of the connections found by *leverage*$_{\mathrm{FDSM}}$ are non-obvious, e.g., the one between "Monk" and "The Dead Zone". But some of the non-trivial components reveal very interesting connections between series: for example, some characters of the two series "Angel" and "Buffy" make an appearance in both series, the same is true for "Xena" and "Hercules", and these series are very tightly connected. Another component connects the series "CSi", "CSI: Miami", "24", "Alias", and "The Shield', all well-known crime series. The series "Queer as Folk" and "The L Word" are both series with homosexual characters and focusing on problems in their relationships, and they can be found in one component as well. The series "Curb your Enthusiasm" is a half-autobiographic series centering on Larry David, who is one of the co-creators of the series "Seinfeld", and both are in the same component.

In summary, the restriction on a set of films for which ground truth can be defined shows strong evidence that using the topmost ranked films by *leverage*$_{\mathrm{FDSM}}$ value as neighbors in an *OMP* will result in a graph in which most edges connect similar films. Not all applications might be best served by using *leverage*$_{\mathrm{FDSM}}$; as we have shown above, also *lift* can be generalized to the fixed degree sequence model (*FDSM*), and other interestingness measures should be easy to adjust to the new null-model *FDSM*. Future research will then need to show when which interestingness measure is best as a basis for a helpful OMP.

Table 9 For each part in the series 'Sex and the City' we show the top five ranked other films according to $\mathit{leverage}_{\mathrm{FDSM}}$ (upper row) and $\mathit{leverage}_{\mathrm{SIM}}$ (lower row)

Series: Sex and the City	1st	2nd	3rd	4th	5th
Season 1	Sex and the City: Season 2	Sex and the City: Season 3	Sex and the City: Season 4	Sex and the City: Season 5	Sex and the City: Season 6: Part 1
	Sex and the City: Season 2	Sex and the City: Season 3	Pretty Woman	American Beauty	Pirates of the Caribbean I
Season 2	Sex and the City: Season 3	Sex and the City: Season 1	Sex and the City: Season 4	Sex and the City: Season 5	Sex and the City: Season 6: Part 1
	Sex and the City: Season 1	Sex and the City: Season 3	Pretty Woman	Miss Congeniality	Forrest Gump
Season 3	Sex and the City: Season 2	Sex and the City: Season 4	Sex and the City: Season 1	Sex and the City: Season 5	Sex and the City: Season 6: Part 1
	Sex and the City: Season 2	Sex and the City: Season 1	Sex and the City: Season 4	Pretty Woman	Sex and the City: Season 5
Season 4	Sex and the City: Season 3	Sex and the City: Season 5	Sex and the City: Season 2	Sex and the City: Season 6: Part 1	Sex and the City: Season 1
	Sex and the City: Season 3	Sex and the City: Season 2	Sex and the City: Season 5	Sex and the City: Season 1	Sex and the City: Season 6: Part 1
Season 5	Sex and the City: Season 6: Part 1	Sex and the City: Season 4	Sex and the City: Season 3	Sex and the City: Season 6: Part 2	Sex and the City: Season 2
	Sex and the City: Season 6: Part 1	Sex and the City: Season 4	Sex and the City: Season 3	Sex and the City: Season 2	Sex and the City: Season 1
Season 6: Part 1	Sex and the City: Season 5	Sex and the City: Season 6: Part 2	Sex and the City: Season 4	Sex and the City: Season 3	Sex and the City: Season 2
	Sex and the City: Season 5	Sex and the City: Season 4	Sex and the City: Season 2	Sex and the City: Season 3	Sex and the City: Season 6: Part 2
Season 6: Part 2	Sex and the City: Season 6: Part 1	Sex and the City: Season 5	Sex and the City: Season 4	Sex and the City: Season 3	Sex and the City: Season 2
	Sex and the City: Season 6: Part 1	Sex and the City: Season 5	Sex and the City: Season 4	Sex and the City: Season 2	Sex and the City: Season 3

7 Summary and open questions

The systematic approach to one-mode projections of bipartite graphs, which was proposed in this paper, can be adjusted to different tasks, mainly by choosing different motifs or by choosing different random graph models as a null-model. There are many open questions regarding when to use which motif and which random graph model. Another open question is how the network analytic approach can help to find more complex association rules. In the following we will discuss these questions before we give a summary of the article.

7.1 Motifs

In ongoing work, we are working on bipartite graphs from a bioinformatic data set. In these bipartite graphs, edges have a binary quality, represented as 'red' or 'green'; there are no multiple edges regardless of the color, i.e., node v is either not connected to node a on the other side, or by a red edge, or by a green edge. In these graphs it makes sense to not only look for the motif *"common neighbor"* but also to evaluate whether this common neighbor is incident to edges of the same color or different colors. Thus, we defined three new motifs: for any given pair of nodes v, w, M_1 counts the number of common neighbors a connected by red edges to v and w, M_2 counts the ones connected by green edges to v and w, and M_3 counts the ones connected by one red to v and one green edge w. Note, that the latter motif is now asymmetric and in the resulting one-mode projection, we introduced *directed edges* to indicate whether v was connected by a red edge to a which was connected by a green edge to w or vice versa. Our first results are very encouraging and reveal clusters of nodes that are either connected by only red or only green edges, indicating groups of molecules that act coherently.

Based on the general framework proposed in this paper, it is now easy to define motifs for general bipartite graphs

Table 10 For each part in the series ‘Buffy the Vampire Slayer’ we show the top five ranked other films according to *leverage*$_{\mathrm{FDSM}}$ (upper row) and *leverage*$_{\mathrm{SIM}}$ (lower row)

Series: Buffy the Vampire Slayer	1st	2nd	3rd	4th	5th
Season 1	Buffy the Vampire Slayer: Season 2	Buffy the Vampire Slayer: Season 3	Buffy the Vampire Slayer: Season 4	Buffy the Vampire Slayer: Season 5	Buffy the Vampire Slayer: Season 6
	Pirates of the Caribbean I	Lord of the Rings: The Fellowship of the Ring	The Matrix	Lord of the Rings: The Two Towers	Independence Day
Season 2	Buffy the Vampire Slayer: Season 3	Buffy the Vampire Slayer: Season 1	Buffy the Vampire Slayer: Season 4	Buffy the Vampire Slayer: Season 5	Buffy the Vampire Slayer: Season 6
	Buffy the Vampire Slayer: Season 3	Buffy the Vampire Slayer: Season 1	Pirates of the Caribbean I	Buffy the Vampire Slayer: Season 4	The Matrix
Season 3	Buffy the Vampire Slayer: Season 2	Buffy the Vampire Slayer: Season 4	Buffy the Vampire Slayer: Season 1	Buffy the Vampire Slayer: Season 5	Buffy the Vampire Slayer: Season 6
	Buffy the Vampire Slayer: Season 2	Buffy the Vampire Slayer: Season 1	Pirates of the Caribbean I	Buffy the Vampire Slayer: Season 4	Spider-Man
Season 4	Buffy the Vampire Slayer: Season 3	Buffy the Vampire Slayer: Season 2	Buffy the Vampire Slayer: Season 5	Buffy the Vampire Slayer: Season 1	Buffy the Vampire Slayer: Season 6
	Buffy the Vampire Slayer: Season 3	Buffy the Vampire Slayer: Season 2	Buffy the Vampire Slayer: Season 1	Buffy the Vampire Slayer: Season 5	Pirates of the Caribbean I
Season 5	Buffy the Vampire Slayer: Season 4	Buffy the Vampire Slayer: Season 3	Buffy the Vampire Slayer: Season 6	Buffy the Vampire Slayer: Season 2	Buffy the Vampire Slayer: Season 7
	Buffy the Vampire Slayer: Season 4	Buffy the Vampire Slayer: Season 3	Buffy the Vampire Slayer: Season 2	Buffy the Vampire Slayer: Season 6	Buffy the Vampire Slayer: Season 1
Season 6	Buffy the Vampire Slayer: Season 5	Buffy the Vampire Slayer: Season 4	Buffy the Vampire Slayer: Season 3	Buffy the Vampire Slayer: Season 7	Buffy the Vampire Slayer: Season 2
	Buffy the Vampire Slayer: Season 5	Buffy the Vampire Slayer: Season 4	Buffy the Vampire Slayer: Season 3	Buffy the Vampire Slayer: Season 2	Buffy the Vampire Slayer: Season 1
Season 7	Buffy the Vampire Slayer: Season 5	Buffy the Vampire Slayer: Season 6	Buffy the Vampire Slayer: Season 2	Buffy the Vampire Slayer: Season 4	Buffy the Vampire Slayer: Season 3
	Buffy the Vampire Slayer: Season 5	Buffy the Vampire Slayer: Season 2	Buffy the Vampire Slayer: Season 3	Buffy the Vampire Slayer: Season 4	Buffy the Vampire Slayer: Season 6

with edges from up to k categories, i.e., weights with k different values. It is, however, still an open question how to deal with weighted edges in general in the proposed framework. This is mainly because it is hard to think of a suitable corresponding random graph model that can deal with the weights.

7.2 Random graph models

There are in general many open questions on when which random graph model is the best suitable. We will now discuss why even more involved random graph models might be applicable: The idea of a random graph model is to model possible outcomes of a network generating process between the nodes. In a market basket analysis, it might thus model the process of a user entering a store and selecting a product. Her decision might depend on her mood, on the products she has already bought, the weather, the popularity of a product, and her general shopping behavior. The question is now, which of these features have to be modeled in order to get a good understanding of which products are often bought together. Of course, the model should not be based on what the user has already bought because this is exactly the correlation we want to understand from the data. In a first approximation, the model should also neglect things like the mood or the weather because it is likely that these effects will smooth out over large data sets. However, an important thing to model might be her general shopping behavior, i.e., how many different products she buys on average per time interval. Regarding films, the Netflix data set shows very clearly that there are very different users: some that watch (or rate) on average more than four films per day and those that hardly rate four films per year (see Fig. 1). It is a

Table 11 For each part in the series 'Star Trek: The Next Generation' we show the top five ranked other films according to *leverage*$_{\mathrm{FDSM}}$ (upper row) and *leverage*$_{\mathrm{SIM}}$ (lower row)

Series: Star Trek: The Next Generation	1st	2nd	3rd	4th	5th
Season 1	Star Trek: The Next Generation: Season 3	Star Trek: The Next Generation: Season 2	Star Trek: The Next Generation: Season 4	Star Trek: The Next Generation: Season 5	Star Trek: The Next Generation: Season 7
	Independence Day	The Matrix	Pirates of the Caribbean I	Lord of the Rings: The Fellowship of the Ring	Lord of the Rings: The Two Towers
Season 2	Star Trek: The Next Generation: Season 4	Star Trek: The Next Generation: Season 3	Star Trek: The Next Generation: Season 5	Star Trek: The Next Generation: Season 7	Star Trek: The Next Generation: Season 6
	Independence Day	The Matrix	Star Trek: The Next Generation: Season 3	Pirates of the Caribbean I	Men in Black
Season 3	Star Trek: The Next Generation: Season 1	Star Trek: The Next Generation: Season 4	Star Trek: The Next Generation: Season 5	Star Trek: The Next Generation: Season 2	Star Trek: The Next Generation: Season 6
	The Matrix	Independence Day	Lord of the Rings: The Fellowship of the Ring	Pirates of the Caribbean I	Spider-Man
Season 4	Star Trek: The Next Generation: Season 6	Star Trek: The Next Generation: Season 5	Star Trek: The Next Generation: Season 3	Star Trek: The Next Generation: Season 7	Star Trek: The Next Generation: Season 2
	Star Trek: The Next Generation: Season 6	Independence Day	Pirates of the Caribbean I	The Matrix	Star Trek: The Next Generation: Season 5
Season 5	Star Trek: The Next Generation: Season 7	Star Trek: The Next Generation: Season 6	Star Trek: The Next Generation: Season 4	Star Trek: The Next Generation: Season 3	Star Trek: The Next Generation: Season 2
	Indiana Jones and the Last Crusade	Raiders of the Lost Ark	The Matrix	Star Trek: The Next Generation: Season 7	Star Trek: The Next Generation: Season 6
Season 6	Star Trek: The Next Generation: Season 5	Star Trek: The Next Generation: Season 7	Star Trek: The Next Generation: Season 4	Star Trek: The Next Generation: Season 3	Star Trek: The Next Generation: Season 2
	The Matrix	Indiana Jones and the Last Crusade	Pirates of the Caribbean I	Independence Day	Lord of the Rings: The Fellowship of the Ring
Season 7	Star Trek: The Next Generation: Season 5	Star Trek: The Next Generation: Season 6	Star Trek: The Next Generation: Season 4	Star Trek: The Next Generation: Season 3	Star Trek: The Next Generation: Season 2
	Star Trek: The Next Generation: Season 5	Indiana Jones and the Last Crusade	Star Trek: The Next Generation: Season 6	Raiders of the Lost Ark	The Matrix

consistent finding that many relationships show this *long-tail behavior* in which a few nodes have a very high number of relations and most have only a few number (Dorogovtsev and Mendes 2003).

If this is the only information we have about a user-product relationship, the FDSM is clearly the model that should be used. With the Netflix data set we had even more information, namely, when a product (film) was introduced into the market and the data at which a rating occurred. Our method gave, e.g., intuitively quite irrelevant recommendations for the film 'Good morning, Vietnam' from 1987, since the Netflix data set was compiled from customer info obtained from 2000 to 2005. Of course, most customers in 2000 can be assumed to already know the film. A more involved random graph model could take such information into account, and, e.g., only allow edges (a rating) for films that are already on the market. In other settings, e.g., biology, a total randomization might not model the systems close enough. Rather, certain nodes might only related to nodes from a subset of the other sides. All of these modifications can easily be included into a corresponding random graph model $\mathcal{G}$. Leverage and lift can then be computed with respect to the newly constructed random graph model. However, it is necessary to show that there is a uniform sampling method from $\mathcal{G}$ and that the expected co-occurrence in this model is monotonic as required by Piatetsky-Shapiro.

7.3 Association rules

We have shown above that the applicability of the simple statistic independence model (SIM) and its corresponding simple bipartite random graph model (SiBiRaG) are limited in assessing the significance of co-occurrence

Table 12 For each part in the series 'Star Trek: Deep Space Nine' we show the top five ranked other films according to *leverage*$_{\mathrm{FDSM}}$ (upper row) and *leverage*$_{\mathrm{SIM}}$ (lower row)

Series: Star Trek: Deep Space Nine	1st	2nd	3rd	4th	5th
Season 1	Star Trek: Deep Space Nine: Season 2	Star Trek: Deep Space Nine: Season 4	Star Trek: Deep Space Nine: Season 6	Star Trek: Deep Space Nine: Season 7	Star Trek: Deep Space Nine: Season 3
	Independence Day	Lord of the Rings: The Fellowship of the Ring	Indiana Jones and the Last Crusade	The Matrix	Pirates of the Caribbean I
Season 2	Star Trek: Deep Space Nine: Season 1	Star Trek: Deep Space Nine: Season 4	Star Trek: Deep Space Nine: Season 3	Star Trek: Deep Space Nine: Season 6	Star Trek: Deep Space Nine: Season 7
	Star Trek: Deep Space Nine: Season 1	Indiana Jones and the Last Crusade	Independence Day	Raiders of the Lost Ark	The Matrix
Season 3	Star Trek: Deep Space Nine: Season 2	Star Trek: Deep Space Nine: Season 4	Star Trek: Deep Space Nine: Season 1	Star Trek: Deep Space Nine: Season 6	Star Trek: Deep Space Nine: Season 7
	Pirates of the Caribbean I	The Matrix	Independence Day	Star Trek: Deep Space Nine: Season 2	Lord of the Rings: The Two Towers
Season 4	Star Trek: Deep Space Nine: Season 2	Star Trek: Deep Space Nine: Season 1	Star Trek: Deep Space Nine: Season 6	Star Trek: Deep Space Nine: Season 3	Star Trek: Deep Space Nine: Season 5
	Indiana Jones and the Last Crusade	Independence Day	The Matrix	Star Trek: Deep Space Nine: Season 2	Star Trek: Deep Space Nine: Season 1
Season 5	Star Trek: Deep Space Nine: Season 6	Star Trek: Deep Space Nine: Season 7	Star Trek: Deep Space Nine: Season 4	Star Trek: Deep Space Nine: Season 1	Star Trek: Deep Space Nine: Season 2
	Indiana Jones and the Last Crusade	Star Trek: Deep Space Nine: Season 6	Raiders of the Lost Ark	Star Trek: Deep Space Nine: Season 7	The Matrix
Season 6	Star Trek: Deep Space Nine: Season 7	Star Trek: Deep Space Nine: Season 5	Star Trek: Deep Space Nine: Season 1	Star Trek: Deep Space Nine: Season 4	Star Trek: Deep Space Nine: Season 2
	Indiana Jones and the Last Crusade	Independence Day	Star Wars: Episode V: The Empire Strikes Back	Star Trek: Deep Space Nine: Season 7	Lord of the Rings: The Two Towers
Season 7	Star Trek: Deep Space Nine: Season 6	Star Trek: Deep Space Nine: Season 5	Star Trek: Deep Space Nine: Season 1	Star Trek: Deep Space Nine: Season 2	Star Trek: Deep Space Nine: Season 4
	Indiana Jones and the Last Crusade	Raiders of the Lost Ark	Star Trek: Deep Space Nine: Season 6	Independence Day	Lord of the Rings: The Two Towers

motifs. This failure can be now explained from both perspectives: the association rules perspective and the network analytic perspective. The statistical model assumes that the probabilities of co-occurrence are independent of all but the degrees of the concerned products. From this assumption it follows that the expected co-occurrence of a given pair of products can be expressed by the product of the single probabilities. However, any user that buys k products induces $\binom{k}{2}$ co-occurrence events. These are of course not independent. Since in most real-world data sets there are some customers which buy many different products $k \gg 1$, SIM cannot even be

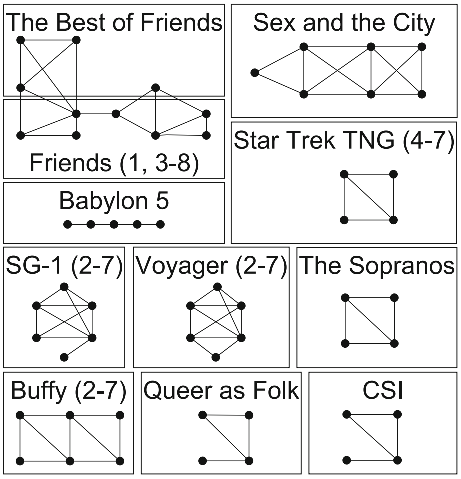


Fig. 12 Each film in a series as defined in the text was connected to its top ten ranked films regarding *leverage*$_{\text{SIM}}$ if the edge was reciprocal. The figure shows all connected components with at least four nodes, and indicates which part from which series can be found in the components. Numbers in *brackets* give the seasons included in the respective component. Abbreviations: "*Star Trek: TNG*" is "The next generation", and "*SG-1*" denotes "Star Gate SG-1"

used to approximate the expected number of co-occurrence events in an null-model for independence. From the network analytic perspective, SiBiRaG fails because it does not maintain the degree sequence of the right-hand side which is essentially responsible for the total number of co-occurrence events.

The question is now how these findings can help to find better association rules. Our first proposal above was to generalize interesting measures to allow for more involved random graph models. Here we want to sketch a generalized approach to find association rules between more than two single products, as it was also discussed by Raeder and Chawla (2011).

Finding good candidates for association rules is not an easy task and has been discussed extensively in literature (Agrawal et al. 1993). It is in general more important to find sensible candidates than to find all possible ones. Thus, a simple idea is to use the one-mode projection onto the products as a basis. A reasonable way would be to choose, e.g., a z-score of 3.29, i.e., a p-value of 0.001 and to connect all pairs of products whose co-occurrence is above this threshold. In this graph, any kind of clustering algorithm can be used to find groups of nodes that are densely interconnected, e.g., the clustering algorithm by Girvan and Newman (2002), a hierarchical clustering (Ward 1963) or the one by Palla et al. (2005) that allows for overlapping cliques. Any of these groups constitutes a set of nodes in which many of the pairs are already significantly co-occurring. This makes them good candidates for co-occurring together. Raeder and Chawla (2011) have proposed this general framework and applied it with success, using a classic OMP as a first step. In future research we propose to use the methods discussed here instead.

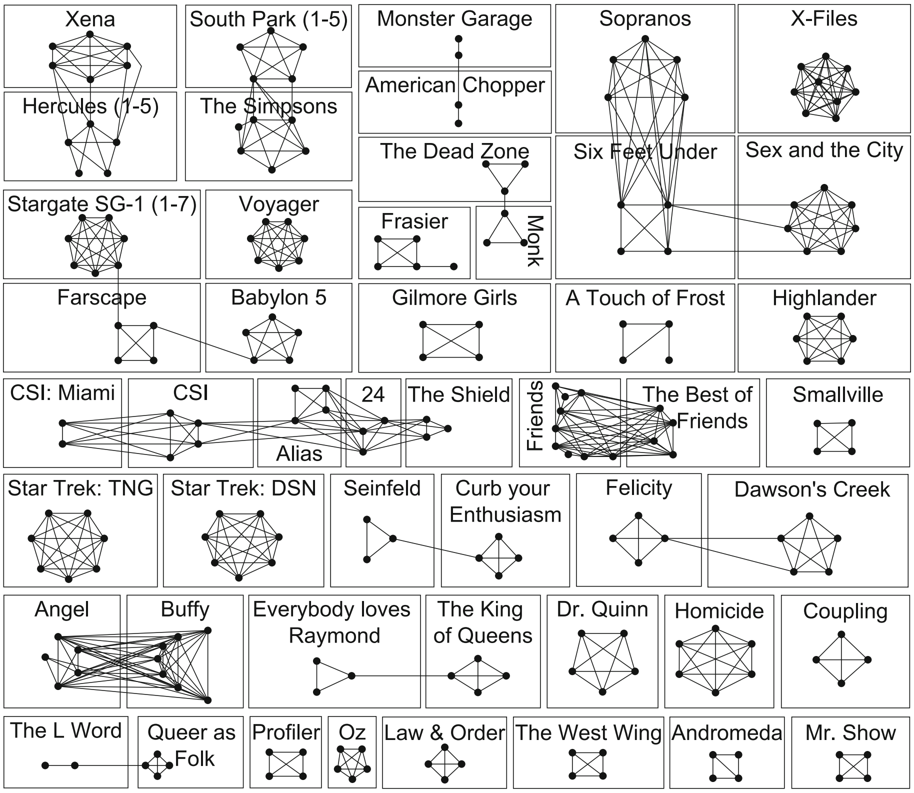


Fig. 13 Each film in a series as defined in the text was connected to its top ten ranked films regarding *leverage*$_{\text{FDSM}}$ if the edge was reciprocal. The figure shows all connected components with at least four nodes, and indicates which part from which series can be found in the components. Numbers in *brackets* give the seasons included in the respective component. Abbreviations: "*Star Trek: TNG*" denotes "The next generation", "*Star Trek: DSN*" denotes "Star Trek: Deep Space Nine", and "*Voyager*" denotes "Star Trek: Voyager"

7.4 Summary

In this article we have proposed a new way to assess one-mode projections of bipartite networks: we have shown that basically each one-mode projection is based on the number of occurrences of a certain motif, namely the *co-occurrence* of two nodes v, w, i.e., the number of their common neighbors. In a second step, the significance of the occurrence of this motif is evaluated. In a classic one-mode projection, a single common neighbor is seen as significant enough to result in an edge between v and w. By drawing a connection between *association rules* and *one-mode projections* we have identified a rich set of significance evaluating rules, the so-called *interestingness measures*. Many of these measures compare the observed occurrence with the expected number under a certain expectation model. We have shown that the classically used expectation model cannot be used for most real-world network data since these show strongly skewed degree distributions on both sides of the graph. We have then argued that the new model, FDSM, needs to be used to maintain the degree distributions on both sides. With this, the expected co-occurrence of any two nodes can be computed and compared with the observed value. Based on two classic interestingness measures, we have then introduced two new similarity measures between any two nodes on one side: $\mathit{leverage}_{\mathrm{FDSM}}$ and $\mathit{lift}_{\mathrm{FDSM}}$. These similarity measures can now be used to build a sparse one-mode projection of the bipartite graph as shown on the example of the $\mathit{leverage}_{\mathrm{FDSM}}$. We have shown that the pairs of nodes with globally highest modified $\mathit{leverage}_{\mathrm{FDSM}}$ and the local lists that rank for each node the other nodes with highest modified $\mathit{leverage}_{\mathrm{FDSM}}$, are either statistically stable with respect to different data samples or they are (almost) empty.

7.5 Outlook

Bipartite graphs belong to the large group of networks where multiple types of actors are connected by one or more relationships, which we call *Multiple-Actor/Multiple-Relationship networks* (MAMuR networks). Huge communities in Web 2.0 applications and the numerous large-scale projects in biology have now created enormous amounts of data that can very often be combined into MAMuR networks. For example, in biology, proteins can interact directly with each other but they are also controlled by other types of molecules, like hormones or miRNA. The past decade has mostly seen analyses of networks representing a single relationship between a single type of actor, e.g., analysis of the protein–protein interaction network on various levels (Milo et al. 2002; Palla et al. 2005; Vázquez et al. 2002). We call these networks *Single-Actor/Single Relationship networks* (SASiR networks). To really understand the behavior of complex networks, it will be inevitable to combine these networks into a MAMuR framework, e.g., to combine into one graph the relation among proteins *and* between proteins and miRNA. The proposed approach to one-mode projection of bipartite graphs is only a first step to condense the information in these graphs into the better understood SASiR networks. But since the framework itself is based on how to evaluate motifs in multipartite graphs, it can also be easily extended to a general analysis of MAMuR networks in various ways and thus we hope to contribute to the understanding of this important and abundant type of networks.

References

Abdi H (2007) The Kendall rank correlation coefficient. In: Encyclopedia of measurement and statistics. Sage, Thousand Oaks

Admiraal R, Handcock MS (2008) Networksis: a package to simulate bipartite graphs with fixed marginals through sequential importance sampling. J Stat Softw 24(8):1–21

Agrawal R, Imielinski T, Swami A (1993) Mining association rules between sets of items in large databases. In: Proceedings of the ACM SIGMOD international conference on management of data 1993, pp 207–216

Alon U (2006) An introduction to systems biology: design principles of biological circuits. Chapman & Hall/CRC

Artzy-Randrup Y, Fleishman SJ, Ben-Tal N, Stone L (2004) Comment on “Network motifs: simple building blocks of complex networks” and “Superfamilies of evolved and designed networks”. Science 305:1107c

Barvinok A (2008) Enumerateing contingency tables via random permanents. Combin Probab Comput 17(1):1–19

Bollobás B (2001) Random graphs, 2nd edn. In: Cambridge studies in advanced mathematics, vol 73. Cambridge University Press, London

Brin S, Motwani R, Ullman JD, Tsur S (1997) Dynamic itemset counting and implication rules for market basket data. In: Proceedings ACM SIGMOD international conference on management of data 1997, pp 255–264

Brualdi RA (1980) Matrices of zeros and ones with fixed row and column vectors. Linear Algebra Appl 33:159–231

Brualdi RA (2006) Algorithms for constructing (0,1)-matrices with prescribed row and column sum vectors. Discrete Math 306:3054–3062

Chen Y, Diaconis P, Holmes SP, Liu JS (2005) Sequential monte carlo methods for statistical analysis of tables. J Am Stat Assoc 100(469):109–120

Cobb GW, Chen YP (2003) An application of Markov chain Monte Carlo to community ecology. Am Math Mon 110:265–288

Dorogovtsev SN, Mendes JF (2003) Evolution of networks. Oxford University Press

Gale D (1957) A theorem on flows in networks. Pac J Math 7:1073–1082

Gionis A, Mannila H, Mielikäinen T, Tsaparas P (2007) Assessing data mining results via swap randomization. ACM Trans Knowl Discov Data 1(3):article no. 14

Girvan M, Newman ME (2002) Community structure in social and biological networks. Proc Natl Acad Sci 99:7821–7826

Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL (2007) The human disease network. Proc Natl Acad Sci 104:8685–8690

Greenhill C, McKay BD (2008) Asymptotic enumeration of sparse nonnegative integer matrices with specified row and column sums. Adv Appl Math 41(4):459–481

Hipp J, Güntzer U, Nakhaeizadeh G (2000) Algorithms for association rule mining—a general survey and comparison. SIGKDD Explor 2(2):1–58

Holmes RB, Jones LK (1986) On uniform generation of two-way tables with fixed margins and the conditional volume test of Diaconis and Efron. Ann Stat 24(1):64–68

Kendall M (1938) A new measure of rank correlation. Biometrika 30:81–89

Li M, Fan Y, Chen J, Gao L, Di Z, Wu J (2005) Weighted networks of scientific communication: the measurement and topological role of weight. Phys A 350:643–656

Ford LR, Fulkerson DR (1962) Flows in networks. Princeton University Press

Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network motifs: Simple building blocks of complex networks. Science 298:824–827

Newman ME (2001a) Scientific collaboration networks I. Phys Rev E 64:016,131

Newman ME (2001b) Scientific collaboration networks II: shortest paths, weighted networks, and centrality. Phys Rev E 64:016,132

Newson R (2006) Efficient calculation of jackknife confidence intervals for rank statistics. J Stat Softw 15(1):1–10

Newman ME, Barabási AL, Watts DJ (eds) (2006) The structure and dynamics of networks. Princeton University Press, Princeton

Palla G, Derényi I, Farkas I, Vicsek T (2005) Uncovering the overlapping community structure of complex networks in nature and society. Nature 435:814–818

Piatetsky-Shapiro G (1991) Knowledge discovery in databases. Discovery, analysis, and presentation of strong rules. AAAI/MIT Press, pp 229–248

Ravasz E, Somera A, Mongru D, Oltvai Z, Barabási AL (2002) Hierarchical organization of modularity in metabolic networks. Science 297:1551–1553

Raeder T, Chawla NV (2011) Market basket analysis with networks. Soc Netw Anal Min 1

Ryser H (1963) Combinatorial mathematics. In: Carus mathematical monograph, vol 14. Mathematical Association of America, Washington

Vázquez A, Flammini A, Maritan A, Vespignani A (2002) Modeling of protein interaction networks. ComPlexUs 1:38–44

Ward JH (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58:236–244

Wasserman S, Faust K (1999) Social network analysis—methods and applications, revised, reprinted edn. Cambridge University Press, Cambridge

Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393:440–442

Zhou T, Ren J, Medo M, Zhang YC (2007) Bipartite network projection and personal recommendation. Phys Rev E 76:046,115

Zweig KA (2010) How to forget the second side of the story: a new method for the one-mode projection of bipartite graphs. In: Proceedings of the 2010 international conference on advances in social networks analysis and mining ASONAM 2010, pp 200–207