

"This material is presented to ensure timely dissemination of scholarly and technical work. Copyright and all rights therein are retained by authors or by other copyright holders. All persons copying this information are expected to adhere to the terms and constraints invoked by each author's copyright. In most cases, these works may not be reposted without the explicit permission of the copyright holder."

The article was published as:

K. Zweig

"How to forget the second side of the story - A new method for the one-mode projection of bipartite graphs" in Proceedings of the International Conference on Advances in Social Network Analysis and Mining, Odense, Denmark, (ASONAM'10), pages 200-207, 2010 (IEEE Computer Society)

"©2010 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE."

How to Forget the Second Side of the Story: A New Method for the One-Mode Projection of Bipartite Graphs

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

Abstract—Many relationships naturally come in a bipartite setting: authors that write articles, proteins that interact with genes, or customers that buy, rent or rate products. Often we are interested in the clustering behavior of one side of the graph, i.e., in finding groups of similar articles or products. To find these clusters, a one-mode projection is classically applied, which results in a normal graph that can then be clustered by various methods. For data with strongly skewed degree distributions, a classical one-mode projection leads to very dense graphs with little information. In this article we propose a new method for a meaningful one-mode projection of any kind of bipartite graph B to a sparse general graph G , using a modified version of the so-called *leverage*. We provide ample experimental evidence that the method creates edges in G only between statistically significant neighbors and that the results are reliable and stable. For this, we present an output sensitive algorithm to compute Kendall’s τ . Moreover, for a subset of films in the Netflix prize data set, we can prove that the proposed method not only detects the statistically significantly co-rented films in the data set but that these are also the films that are the most similar ones by content. Thus, our method cannot only be used for the one-mode projection of bipartite graphs in general but also especially for any kind of market basket data to find pairs of most similar products as needed for, e.g., recommendation systems.

I. INTRODUCTION

The clustering of graphs is one of the most useful theoretical approaches in the field of network analysis. It helps to reduce information, to predict the behavior of less well-known objects by grouping them with better known objects and to find groups of similar objects, e.g., for recommendation systems. For general graphs, there are many methods of clustering that have been successfully applied to various data sets [4], [8]. But many data sets come in the form of so-called bipartite graphs, i.e., they describe the relationship between objects of different types. A famous data set in this realm is the Netflix competition data set, consisting of 100 million ratings of films given by users. Each rating is determined by one user ID, a film ID, and the rating (a number between 1 and 5). The Netflix competition asked to build a recommendation system that beats the company’s own system. A first approach is to find which films are similar to each other, by using the information of how often two films were rented (and rated) by the same user.

The idea behind this approach is that if many users view both films, these films might be similar to each other.

The most common way to deal with such a bipartite graph of which one side is of main interest, is to turn it into a general graph by a so-called ‘one-mode projection’. For this, two objects from the same side are connected if they share neighbors on the other side. The easiest approach is to connect them if they share at least one neighbor. Concerning the Netflix data set, this implies that if a user has rented k films in total, this will cause all k films to be connected to each other. Unfortunately, the degree distribution of the users has a very long tail where one user saw almost all films, while a second user saw approximately 14,000 out of 17,770 films. A classical one-mode projection would thus result in a network of films where almost every film is connected to every other film. Another approach is to weight the edges by the number of times they have been co-rented, either directly or by a more complicated scheme as proposed by Newman [11]: for each article that two authors wrote together in a group of k authors in total, the weight of the link between them is increased by $\frac{1}{k-1}$.

This still leaves us with the problem that the weights can have very different meanings if the objects have a degree distribution with a long tail. This is the case for most real-world data, i.e., some authors write a lot of papers while most write only a few, and some films are rented by 50% of all users while most are rented by only a few percent. It is thus more significant that the film “VeggieTales: Duke and the Great Pie War”, which itself was only rented 40 times in a set of 10,000 customers, is co-rented 23 times together with “VeggieTales: Lyle the Kindly Viking” (itself rented 73 times), while even 1197 co-rentings of the films “Pretty Woman” (4080 rentals) and “Star Wars: Episode V: The Empire Strikes Back” (1930 rentals) are insignificant because both films are so highly popular. Thus, the question is: given the degrees of two objects on one side of a bipartite graph and the number of shared neighbors on the other side, called their *co-occurrence*, is it significant? And furthermore: if, given one object o_1 , we rank the other objects o_2 by the significance of their co-occurrence, will the most highly ranked objects also be the

ones most similar to o_1 by any intuitive measure? This general question is not only of interest for films and customers but for any kind of bipartite data set in which a relationship between elements from the two sets can occur at most once. Examples for other data sets of this form are proteins and their biological function, authors and articles [11], Darwinian finches and their preferred islands [5], or ratings from customers of products in a forum.

In this article we use a classical network analytic approach to assess the significance of a given co-occurrence by comparing it with its expected value in a suitable random graph model. The article shows how to evaluate this information to define a sparse one-mode projection with only $O(n)$ edges from any bipartite graph. As a side result we present an $O(n + out)$ algorithm for computing Kendall's τ , a *rank correlation coefficient*, which significantly improves the runtime compared with the classical $O(n \log n)$ algorithm in those data sets in which the rankings are expected to be almost the same.

The paper is organized as follows: Section II introduces basic terms and models. Section III defines the random graph model by which the significance of a given co-occurrence is tested. Section IV applies the method to various samples from the Netflix data set. The findings are then discussed in Section V.

II. DEFINITIONS

A *graph* is composed of a set of nodes V with $n = |V|$ the number of nodes, and a set of edges $E \subseteq V \times V$ with $m = |E|$ the number of edges. If the node set can be partitioned into two sets V_0 and V_1 such that all edges $e = (v, w)$ are between nodes $v \in V_0$ and $w \in V_1$, and none between nodes from the same set V_i , the graph is said to be *bipartite*. In the following, all graphs will only contain simple edges, i.e., no two edges contain the same pair of nodes, and no selfloops. For any two nodes v, w from the same set V_i , we define the *co-occurrence* $coocc(v, w)$ as the number of nodes z in the other set such that (v, z) and $(w, z) \in E$. The *degree* $deg(v)$ of any node is defined as the number of edges it is contained in, i.e., the number of neighbors it has. Given some fixed order of the vertices in V_0 and V_1 , we define the *degree sequence* L and R of the graph as the sequence of the degrees of V_0, V_1 .

In general, any method that turns a bipartite graph $B = (V_0 \cup V_1, E)$ into a graph $G(V_i, E')$ on one of the two nodes set V_i , i.e., a *one-mode projection*, consists of first computing some weight between all pairs of nodes in V_i by defining a *similarity measure* $s : V_i \times V_i \rightarrow R$. The according matrix can be turned into an adjacency matrix in multiple ways, e.g., by creating an edge between all nodes with at least a given *threshold similarity*, by connecting each node with the k neighbors with highest similarity to it, or by simply using the similarity as a weight for a weighted graph. In the following we will propose a new similarity measure for nodes in one side of a bipartite graph and experimentally assess its validity and quality.

III. MODELS AND METHODS

Using a classical network analytic approach for evaluating the significance of a structural graph measure, we compare the co-occurrence of any two films with its *expected value*. Since the film-user scenario resembles data sets from *market basket analysis*, we use the modified version of a so-called *interesting measure* from that area. Market basket analysis tries to deduce rules of the form $X \rightarrow Y$ denoting that if X is bought, there is a significant probability that Y is bought as well. In market basket analysis, the *support* $supp(X)$ of a product or subset of products X is defined as the frequency with which it is bought (together). If X contains only one film v , $supp(X) := deg(v)/r$, where r denotes the number of baskets (users); if X contains two films v, w , $supp(v, w) := coocc(v, w)/r$. The *leverage* is then defined as the difference between $supp(v, w)$ and its expected value. In its original definition by Piatetsky-Shapiro, the expected value was assumed to be given by $supp(v) \cdot supp(w)$ [13]. The underlying simple model assumes that if film v is rented by 30% of all customers and film w by 70%, both films are expected to be rented by 21% if they are unrelated. This simple independence model (SIM) thus assumes that all customers rent each film v with the same probability $deg(v)/r$. In the limit of a large number of films, this leads to a degree distribution of the customers that expectedly follows a Poissonian distribution. Although this sounds like a reasonable random model, it can be quickly seen that it is not suitable for those data sets that have a customer-degree sequence with a long tail (skewed degree distribution) [14]. Since most real-world graphs show a strongly skewed degree distribution [2], SIM should not be used to assess the expected co-occurrence or the expected support in these cases: rather, as discussed by Gionis et al. [7] and theoretically analyzed by Zweig and Kaufmann [14], the expected co-occurrence should be assessed in the *fixed degree sequence model* (FDSM): given a graph $G = (V = \{V_0, V_1\}, E)$ and the corresponding degree sequences L and R , we define $\mathcal{G}(L, R)$ as the set of all bipartite graphs with the same degree sequences (and no multi-edges). The expected co-occurrence (expected support) is then defined as the average co-occurrence (support) over all graphs in $\mathcal{G}(L, R)$. Since the cardinality of this set is enormous even for bipartite graphs of moderate size [3], [6], it is in general not possible to enumerate all graphs from $\mathcal{G}(L, R)$. Fortunately, it is possible to sample from $\mathcal{G}(L, R)$ with a simple random walk procedure: starting from a graph in $\mathcal{G}(L, R)$, in each step two edges (v, w) and (x, y) are drawn uniformly at random from the set of all edges. If a swap of the target nodes of these two edges does not lead to a multi-edge, i.e., if neither (v, y) nor (x, w) is already in E , (v, w) and (x, y) are replaced by (v, y) and (x, w) . Note that this swap operation maintains the degrees of all affected nodes and thus the resulting graph is also in $\mathcal{G}(L, R)$. Note that the original bipartite graph, which is itself in $\mathcal{G}(L, R)$, can be used as a simple starting point for the random walk. After a sufficient number of random walk steps, the procedure is guaranteed to end at each of the graphs

in $\mathcal{G}(L, R)$ with the same probability [5]. Sampling a sufficient number of graphs in this way and averaging over the co-occurrences of interest in them will then give an approximation of the value of the expected co-occurrence in $\mathcal{G}(L, R)$. This approximated value will be denoted by $coocc_{FDSM}(v, w)$. The modified version of the leverage is then given by:

$$leverage_{FDSM}(v, w) = coocc(v, w) - coocc_{FDSM}(v, w). \quad (1)$$

This defines a similarity measure on the set of all pairs of nodes from one side and can be computed by the following steps:

- 1) Based on the given data set and its graph representation $G = (V_0 \cup V_1, E)$, compute $coocc(v, w)$ for each pair of nodes on the side of interest.
- 2) Based on the given data set, compute the degree sequences L and R .
- 3) Sample a large set H of graphs from $\mathcal{G}(L, R)$ in the following way:
 - a) Start from G which itself is obviously in $\mathcal{G}(L, R)$.
 - b) Fix a number of random walk steps to be performed, e.g., $m \log m$.
 - c) Perform this number of random walk steps, where each step consists of choosing two edges at random, check whether they can be swapped and swap them if possible. A non-swappable pair of edges needs to be counted as a random walk step to assure that the random walk stops at each graph with the same probability.
 - d) Let G' be the graph from $\mathcal{G}(L, R)$ resulting from the random walk procedure. For each pair of nodes on the side of interest compute the co-occurrence and store it.
- 4) For each pair of nodes on the side of interest, average over the observed co-occurrences in the sampled random graphs, denoted by $coocc_{FDSM}(v, w)$.
- 5) For each pair of nodes v, w on the side of interest, compute $leverage_{FDSM}(v, w)$. Store triples $v, w, leverage_{FDSM}(v, w)$ in one global list GL or store pairs $w, leverage_{FDSM}(v, w)$ in one list local $LL(v)$ for each node v .
- 6) Sort list(s) non-increasingly by value.

Note that the necessary length of the random walk to guarantee a uniform sampling from $\mathcal{G}(L, R)$ is not known. However, in most real-world data sets there are many nodes with the same degree which provides for an internal check: the observed co-occurrence value $coocc_{FDSM}(v, w)$ only depends on the degrees of v and w . This implies that every pair of nodes (v, w) and (x, y) with the same degrees $deg(v) = deg(x)$ and $deg(w) = deg(y)$ should have the same $coocc_{FDSM}(v, w)$ value. By fixing some node v and plotting $coocc_{FDSM}(v, w)$ against the degree of w on the x-axis all values for films w, w' with $deg(w) = deg(w')$ should fall on top of each other. As long as the variance of these values is too high, either the length of the random walk or the number of sampled graphs has to be increased.

As described above, the proposed modified leverage $leverage_{FDSM}(v, w)$ defines a similarity measure between all pairs of objects on one side of the graph. As with other similarity data, this can be turned into a sparse graph in many ways: It can either 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.

A. 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 films will be co-rented together or which authors are most likely to produce another article together, but also whether the films are similar by content 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 judgement of such a conjecture is that we seldomly have something like a *ground truth* with which we can compare our results. Luckily, the Netflix prize data set [9] 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. There are over 480,000 distinct customers, identified by an ID between 1 and 2,600,000. The degree distributions of customers and films are both highly skewed. For each rating event, the customer ID, the film ID and the rating from 1 to 5 ('very bad' to 'very good') is presented. Additionally, a second file assigns to each film ID the film's title and its publishing year. The data allows for different quality assessment techniques:

- 1) Since the data set is so large, it can be partitioned 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 $GL(DS_i)$. 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 $LL(DS_i, v)$ containing its up to 100 best neighbors w sorted by their modified leverage $leverage_{FDSM}(v, w)$. To ease understanding, in the following the term *leverage* will be used to signify the modified leverage as proposed above and $leverage(v, w)$ is understood as $leverage_{FDSM}(v, w)$ in the *FDSM* model.

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, $GL(DS_i)$ and $LL(DS_i, v)$ 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.

B. 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* [10]. An easier formulation was given by Abdi on which we rely here [1]. 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 $1, 2, 3, \dots, n$, 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 x -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 $[5, 3, 4, 2, 1]$ 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 [12]. 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 $\pi(x) < i$. The detailed proof is omitted due to space restrictions.

The runtime is in $O(n)$ 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 $\Pi = [n, \dots, 3, 2, 1]$, the runtime is in $\Omega(n^2)$. With this rank correlation coefficient, the different global and local rankings in the 20 data samples created from the Netflix data set can now be assessed.

IV. EXPERIMENTAL RESULTS

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.

A. Global rankings

For every 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 possible more than 157,000,000 distinct pairs of films, the best ranked 1000 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 I¹. Interestingly, these pairs give rise to three distinct

¹Note that the order was chosen for displaying reasons - none of the data samples directly showed them in this order.

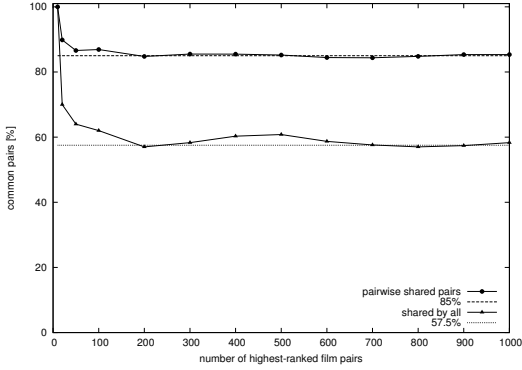
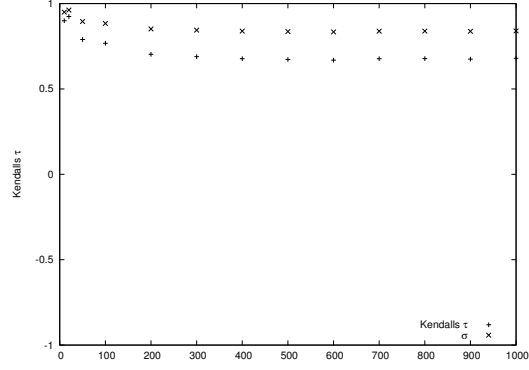


Fig. 1. Average percentage of pairwise shared pairs of films in $GL(DS_i)$ and $GL(DS_j)$, restricted to the first k rankings in all 20 data samples, and percentage of pairs of films listed under the first k rankings in all 20 data samples (maximal consensus).

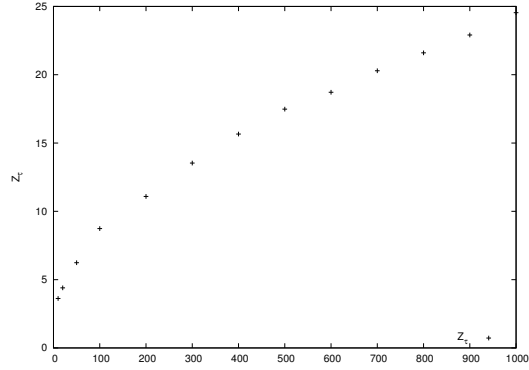
3-cliques and one single pair of films, and all films in the same component are obviously highly related: All Lord-of-the-Rings sequels are connected to each other, once in the normal cut and once in the extended version - but there is not yet a connection between the two components. Similarly, Star Wars Episode *IV* to *VI* are all pairwise related and thus build a triangle. The last pair makes a connection between both volumes of *KillBill*.

If more than the first 10 rankings are considered, the percentage of common pairs drops, as depicted in Fig. 1. 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 , all data samples agree on around 57% pairs of films.

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. Fig. 2(a) 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 $N > 10$ [1]. τ 's significance can thus be tested by computing $Z_\tau = \tau/\sigma_\tau$, which denotes how many standard 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 ($k = 1000$) and is thus highly significant (s. Fig. 2(b)).



(a)



(b)

Fig. 2. 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.

B. 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 $leverage(v, w)$. We restrict them to those w with $leverage(v, w) > 10$, 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 $leverage(v, w) > 10$ 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

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

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

TABLE I

PAIRS OF FILMS WITH THE 10 HIGHEST LEVERAGE VALUES IN ALL 20 DATA SAMPLES. THE LEVERAGE OF ALL PAIRS IS AT LEAST 725 IN ALL DATA SAMPLES.

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 $lev_{max}(v)$ it has with any other film w , the number of neighbors w with at least $leverage(v, w) > 10$ (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 $|lev_{10}(v)|$ (s. Fig. 3(a)). 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 of two films, if it is significant, is a reliable measure that will identify the same significant neighbors in different data sets.

Fig. 3(b) shows the dependence of the maximal leverage of a film $lev_{max}(v)$ and its number of significant neighbors $|lev_{10}(v)|$ and Fig. 3(c) shows the dependence of $lev_{max}(v)$ and the size of the consensus set $|cons(v)|$. It can be seen that if the maximal leverage is below 15, there are mostly less than 10 significant neighbors and never more than 7 neighbors in the consensus set. The diagrams show in general that a small leverage value is correlated with a low number of consensus neighbors. Thus, a low maximal leverage $lev_{max}(v)$ 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. Fig. 3(c) and Fig. 3(d) finally show that for all films whose maximal leverage 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 secondly to give statistically reliable recommendations for those objects that have significant neighbors. Thus, any reasonable method that uses the modified leverage 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.

C. 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 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 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

All of these recommendations seem to be reasonable and some are even interesting and non-obvious. But given the other

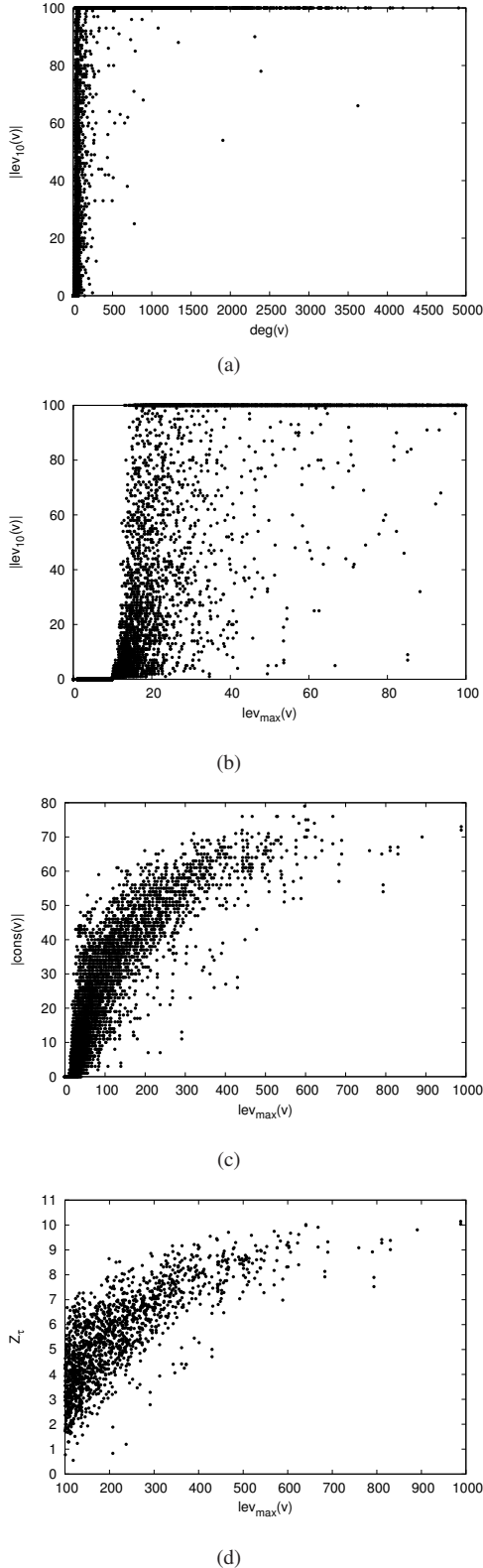


Fig. 3. Assessment of local ranking correlations. (a) Scatterplot of degree $deg(v)$ and number of significant neighbors $|lev_{10}(v)|$. (b) Scatterplot of $lev_{max}(v)$ and $|lev_{10}(v)|$ restricted to $lev_{max}(v) \in [0 : 100]$. (c) Scatterplot of $lev_{max}(v)$ and $|cons(v)|$. (d) Scatterplot of $lev_{max}(v)$ and Z_{τ} restricted to $lev_{max}(v) > 100$.

17,767 films in the data set it is hard to judge whether these are really the *best* recommendations. Luckily, the data set at hand allows for some quality measure in this realm by concentrating on film series. To find them, we have extracted all film titles who had the key word ‘Season’ in it, standing for one part of a series. We kept all series that had at least one volume that was published in 1990 or later (s. Table II for an overview, data on series with less than 3 parts omitted). Given one film x out of a series and its list $LL(x)$ of the first 100 highest-ranked films, we require that the highest-ranked film in $LL(x)$ 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
 - a) the average rank $\overline{first}(S)$ of the first listed sequel from the same series (1,00 is optimal)
 - b) the average rank $\overline{last}(S)$ of the last listed sequel from the same series ($n(S) - 1$ is optimal).

Table II lists all optimal results in green and it can be easily seen that 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.

V. DISCUSSION

In this article we have proposed a new network analytic similarity measure, the modified leverage, that assesses whether two nodes in a bipartite graph share a significant number of neighbors on the other side of the graph. This similarity measure can then be used to build a sparse one-mode projection of the bipartite graph. We have shown that the pairs of nodes with globally highest modified leverage and the local lists that rank for each node the other nodes with highest modified leverage, are either statistically stable with respect to different data samples or they are (almost) empty.

The FDSM implicitly assumes that in a given system of authors and articles, films and customers or any other system from which the bipartite graph stems, the degree distributions on both sides are universal: given any data sample from the same environment, we will see approximately the same degree distribution. If the degree distributions are a given fact, the independence assumption needs to take them into account, by

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... ²	3	0,00	0,00	-	-
Roswell	3	100,00	100,00	1,00	5,67
Russell Simmons... ³	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... ⁴	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... ⁵	5	60,00	40,00	1,00	10,00
A Touch of Frost	6	50,00	66,67	1,50	49,25
Dr. Quinn... ⁶	6	66,67	16,67	2,00	27,00
Frasier	6	100,00	50,00	1,00	32,00
Hercules: The ... ⁷	6	50,00	0,00	-	-
Highlander	6	100,00	100,00	1,00	6,50
Homicide: Life on... ⁸	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... ⁹	7	100,00	100,00	1,00	6,29
Star Trek: The Next... ¹⁰	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

TABLE II

AVERAGE QUALITY ASSESSMENT FOR RECOMMENDATIONS OF ALL SEQUELS IN A GIVEN SERIES S AS DESCRIBED IN THE TEXT WITH RESPECT TO THE FIRST DATA SAMPLE (FIRST 10,000 SUBSEQUENT USER IDS) AS DESCRIBED IN THE TEXT. ALL OPTIMAL VALUES ARE COLORED IN GREEN.

building all possible graphs with the same degree distribution. Of course, the modified leverage of two films then depends on the whole bipartite graph in this model. An important question is thus how large a bipartite graph needs to be and which data to include. E.g., our method gave 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 2003. Of course, most customers in 2000 can be assumed to already know the film, and the implicit assumption in the FDSM that each customer in 2000 could also have chosen this film instead of another film might not be valid. The bipartite graph should thus be restricted such that the implicit assumptions of the FDSM can be assumed to be valid. Future research will further evaluate this question.

ACKNOWLEDGMENT

The author wants to thank Winfried and Lena Z. for their discussions and general support in writing this article.

REFERENCES

- [1] Hervé Abdi. *Encyclopedia of Measurement and Statistics*, chapter The Kendall Rank Correlation Coefficient. Thousand Oaks (CA): Sage, 2007.
- [2] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.
- [3] Richard A. Brualdi. Matrices of zeros and ones with fixed row and column vectors. *Linear Algebra and Its Applications*, 33:159–231, 1980.
- [4] Aaron Clauset. Finding local community structure in networks. *Physical Review E*, 72:026132, 2005.
- [5] George W. Cobb and Yung-Pin Chen. An application of markov chain monte carlo to community ecology. *The American Mathematical Monthly*, 110:265–288, 2003.
- [6] Persi Diaconis and Anil Gangolli. Rectangular arrays with fixed margins. *Institute for Mathematics and Its Applications*, 72:15–41, 1994.
- [7] Aristides Gionis, Heikki Mannila, Taneli Mielikäinen, and Panayiotis Tsaparas. Assessing data mining results via swap randomization. In *KDD'06*, 2006.
- [8] Michelle Girvan and Mark E.J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99:7821–7826, 2002.
- [9] <http://www.netflixprize.com>.
- [10] M. Kendall. A new measure of rank correlation. *Biometrika*, 30:81–89, 1938.
- [11] Mark E.J. Newman. Who is the best connected scientist? A study of scientific coauthorship networks. arXiv: cond-mat/0011144, November 2000.
- [12] Roger Newson. Efficient calculation of jackknife confidence intervals for rank statistics. *Journal of Statistical Software*, 15(1):1–10, 2006.
- [13] G. Piatetsky-Shapiro. *Knowledge Discovery in Databases*, chapter Discovery, analysis, and presentation of strong rules, pages 229–248. AAAI/MIT Press, 1991.
- [14] Katharina A. Zweig and Michael Kaufmann. The expected co-occurrence in bipartite graphs with fixed degree distribution. In *submitted*, 2010.