Exploiting Phase Transitions for the Efficient Sampling of the Fixed Degree Sequence Model

Christian Brugger, André Lucas Chinazzo,
Alexandre Flores John,
Christian De Schryver, Norbert Wehn
Microelectronic System Design Research
Group, University of Kaiserslautern, Germany
Email: {brugger, schryver, wehn}@eit.uni-kl.de,
{chinazzo, floresj}@rhrk.uni-kl.de

Andreas Spitz
Graph Theory and Network
Analysis Group, Heidelberg
University, Germany
Email: spitz@informatik.
uni-heidelberg.de

Katharina Anna Zweig
Complex Network Analysis
Group, University of
Kaiserslautern, Germany
Email: zweig@cs.uni-kl.de

***Abstract*—Real-world network data is often very noisy and contains false-positive edges and misses false-negative ones. These superfluous and missing edges can be identified by evaluating the statistical significance of the number of common neighbors of any two nodes. To test whether the number of common neighbors, the *co-occurrence* of two nodes, is statistically significant, the values need to be compared with the expected occurrence in a suitable random graph model. Zweig has proven that for networks with a skewed degree sequence, a random graph model which maintains the degree sequence, the *fixed degree sequence model*, needs to be used instead of estimating the numbers from a simple independence model [1]. However, using that random graph model practically means to sample networks by a Markov chain approach and to measure the occurrence of each subgraph or the co-occurrence of each pair of nodes in each of the samples. Thus, the computational complexity depends on both, the length of the Markov chain and the number of samples. In this article we show, based on ground truth, that there are various phase transition-like tipping points that enable to choose a comparatively low number of samples and that reduce the length of the Markov chains without reducing the quality of the significance test.**

I. Introduction

The identification of so-called *network motifs*, i.e., subgraphs whose occurrence is statistically significant, is of general interest, especially in biological data sets [2]. The significance of a subgraph is tested by counting its occurrence in an observed real-world network and by comparing it to the expected occurrence in a suitable random graph model. A *random graph model* is defined as a set of graphs together with a probability mass function that assigns a probability to each member of the set, summing up to 1. While not specifically a network motif, the number of common neighbors of two nodes x, y, their so-called *co-occurrence* $coocc(x, y)$, can also be tested on its statistical significance in the same way. Zweig and Kaufmann have proven that, while a simple independence model estimates the expected co-occurrence of two nodes to be $deg(x)deg(y)/2m$, where $deg(x)$ denotes the *degree* of node x, this model is wrong if the degree sequence is skewed [1], [3]. The *degree sequence* $DS(G)$ of a graph G is defined as the sequence of degrees of the nodes of G in some fixed order. Instead, a more detailed random graph model needs to be used, such as the set $\mathcal{G}(DS)$ of all simple graphs with the same degree sequence as the observed network and uniform probability to sample any of it—this model is in the following called the *fixed degree-sequence model* or **FDSM**. Note that a graph is *simple* if it does not contain multiple edges between the same nodes and no self-loops.

The statistical significance of the number of common neighbors can be used to do *link assessment*, i.e., to evaluate whether an existing edge in a graph is likely to be a true-positive and whether a non-existing edge is likely to be a false-negative [4], [5], [6]. A *link assessment* results in a ranking of all pairs of nodes, where (existing) edges with a high ranking are assumed to be true-positives and pairs of nodes not yet connected by an edge but with a high ranking are assumed to be false-negatives. This assumption can be quantified for those networks with an assigned *ground truth*, i.e., where there is some noisy edge set containing false-positives and false-negatives, and a verified set of edges.

Unfortunately, there is no fast way of sampling from $\mathcal{G}(DS)$. **ToDo: [Andreas, macht der Satz so Sinn, oder ist das exact sampling scheme auch auf einer MC aufgebaut?]** An *exact sampling scheme* as proposed by Del Genio et al. is practically not useful for large data sets [**?**]. There is, however, a Markov chain with an unknown mixing time. Since in this article we only deal with bipartite graphs, the following describes the Markov chain for sampling from the set of all bipartite graphs with the same degree sequences on both sets of nodes (see, for example, [11]): starting from the observed network, in every step two edges $e_1 = (a, b)$ and $e_2 = (c, d)$ are chosen uniformly at random from all edges and it is tested whether they are swappable, i.e., whether $e'_1 = (a, d)$ and $e'_2 = (c, b)$ are already in the graph. If so, the tested edges e'_1 and e'_2 are inserted into the graph and e_1, e_2 are deleted from the graph. A pre-defined number of swap tests are done—regardless of the result of the test—and it can be shown that after a sufficient number of tests, the resulting graph is any one of $\mathcal{G}(DS)$ with the same probability. The sufficient number, the so-called *mixing time*, is to date unknown for this specific Markov chain—and the known upper bounds (e.g., [**?**], [12]) are practically useless because they are either too big like Jerrum et al's result in $O(n^{14} \log^4 n)$ or not computable for large networks like Brualdi's result that the convergence time depends on the spectral gap of the transition matrix of the Markov Chain [13]—computing the latter would require to know all possible graphs in $\mathcal{G}(DS)$ and their transitions.

Depending on the size of the graph and the resulting data structures to store it in memory, a single swap test and updating the data structure(s) after a successful swap cost between $O(1)$ and $O(\log n)$ or $O(\min\{deg(x), deg(y)\})$. Note that, as for all Markov chains with known degrees, there is the probability of an importance sampling, but again, it is not of practical use for large data sets [8].

A *safe* number of steps is considered to be in $O(m \log m)$ which is the lower bound such that, expectedly, every edge is chosen at least once for a swap test. Often, the safe number of steps is also used for a so-called *burn-in* phase where one tries to get away from the often very strongly structured observed network to one that is more random in its structure. From this instance, a chain of swaps is started, where every xth resulting graph is tested for its structural features—these graphs comprise the *sample* against which the statistical significance of structural features of the observed graph is tested. Again, a "safe" size of this set is often used, for example, $10,000$ samples. **ToDo: [Nina: Do we have a ref here?]** This scheme will be called the *serial burn-in (sampling) scheme* in the following.

To empirically analyze the necessary burn-in length, Gionis et al. proposed some data-mining specific, empirical convergence tests, namely observing the convergence of the number of so-called *frequent item sets* or their frequencies [7], [8]. By plotting the number of frequent item sets in dependence of the number of swap tests, it can be seen that this number stabilizes (Gionis et al., Fig. 4). However, counting frequent item sets is computationally expensive and Gionis et al. do not provide an online stopping criterion that allows to stop sampling.

This article looks at link assessment in bipartite graphs, i.e., given a graph $G = (V_L \cup V_R, E)$ with a node set V_L, a node set V_R and an edge set E only connecting nodes of V_L with nodes of V_R, we assess whether any two nodes in V_R have a statistically significant number of nodes in common—this information can be used to build insightful one-mode projections of bipartite graphs [1], [3], [9]. This article provides two online heuristics, one to determine a sufficient number of the number of swaps, one to determine a sufficient number of samples. For the first time, the quality of the resulting statistics is tested against ground truth for the link assessment task which shows astonishing *tipping points*: while at first—with low numbers of swap tests—the link assessment is not good, a small increase has a strong impact on the quality of the link assessment—an effect that is often called a *phase transition*. Similarly, computing the co-occurrences for all pairs of nodes of interest in one sample, is costly (in $\Omega(n^2)$ to $O(n^3)$, depending on data structures and density). Thus, reducing the number of samples is also an issue. Again, we find that there is a phase-transition-like behavior in the number of samples. Finally, especially for bipartite graphs in which the hidden connections between nodes on one side of the graph are assessed [1], [3], [9], it can make sense to reduce the graph by sampling from nodes of the other side. For example, in market-basket data with millions of customers but only 10,000 of products, it might not be necessary to look at the whole data set but to reduce it to the market baskets of 50,000 customers. Again, we find a phase-transition like behavior that points to an optimal set. Optimizing these parameters, we achieved speedups of up to one order of magnitude.

Our novel contributions are:

- To show that there is a phase-transition like behavior in the length of the burn-in phase, the number of swaps, the number of samples and the sample size of the left-hand side in a bipartite graph.
- We present two online heuristics to estimate just the required #samples and #users.
- We demonstrate their effectiveness and stability in numerical studies for multiple datasets.

The next section introduces definitions from statistics to assess the statistical significance of the co-occurrence of two nodes in a bipartite graph.

II. Definitions

Given a bipartite graph $G = (V_L \cup V_R, E)$ as defined above, the co-occurrence of two nodes $x, y \in V_R$ is defined as the number of their common neighbors in V_L. This value is bound from above by $\min\{deg(x), deg(y)\}$, the minimal degree of both nodes. Thus, its absolute value cannot be used to understand its significance as nodes with a small degree would always be disfavored. Their *expected* co-occurrence with respect to some random graph model, e.g., the $FDSM$ ($coocc_{FDSM}(x, y)$), is defined as the expected co-occurrence in all graphs in the model, given their probability. If it is not possible to compute it, it is approximated by the average observed co-occurrence in a uniform sample from the random graph model. In this article we will identify the two but please bear in mind that the approximation quality is depending on the sample size and on the quality of the sample. Given a sample from a random graph model, two statistical values can be computed with respect to one pair of nodes and their observed co-occurrence: the p-value denoting the fraction of samples observed in which the co-occurrence of x and y was at least as high as the observed one; the z-score of the observed co-occurrence given the co-occurrence distribution of x and y in the sample:

$$p\text{-}value(x, y) = \sum_{i=1}^{\#\text{samples}} \begin{cases} 1, & \text{if } coocc_i(x, y) > coocc(x, y) \\ 0, & \text{otherwise} \end{cases},$$

$$z\text{-}score(x, y) = \frac{leverage(x, y)}{\text{stddev}\left(\{coocc_i(x, y)\}_{i=1,..,|samples|}\right)},$$

Based on previous work by Zweig et al. [1], [3] and Horvat et al. [9], two nodes with a high z-score, or low p-value, are connected in a one-mode projection of the bipartite graph. It often shows that they are also semantically similar, for example this method can identify similar movies [1], [3] or biologically similar proteins [4]. In this article, we rank all edges by their p-value and break ties by the z-score.

III. Ground truth and PPV_k

We are using two data sets, the Netflix competition data set and a medium size MovieLens data set[1]; both data sets show ratings of films by a number of users. By setting a threshold, the data can be represented as a bipartite graph between users and movies, where an edge (u, j) represents that user u likes

[1] The 100k MovieLens data set, available from http://grouplens.org/datasets/movielens/.

film j. By finding significant co-occurrences between any two movies i, j, a one-mode projection can be built [3]. We use a ground truth data set based on movie sequels like Star Wars I to VI, as compiled by Horvat and Andreas Spitz (University Heidelberg), first used in a paper by Horvat and Zweig [6]. This ground truth can be used for both data sets. The idea is that, for a given set of sequels like the Star Wars movies or all James Bond movies, the most significant co-occurrences are assumed to be with other sequels from the same set. Thus, ranking all pairs of films (where at least one is a sequel from a series) by the significance of their co-occcurrence, the most significant pairs should be sequels from the same set. The quality of such a ranking can be evaluated by the *positive predictive value* PPV_k, where the k indicates the number of pairs of films in the ground truth and the PPV is the fraction of correctly identified pairs from the ground truth in the set of the k highest ranked pairs of films. This measure was proposed by Liben-Nowell and Kleinberg as more meaningful in the very unbalanced link prediction problem which is very similar to the problem proposed here [10].

IV. Phase Transitions

For each parameter of the algorithm, i.e., the number of swaps, the number of samples and the length of the burn-in-phase, the PPV_k can be computed for the resulting ranking of the co-occurrences. Here we show that the quality of the ranking improves very suddenly in all of the three parameters. This indicates on the one hand that these parameters are often smaller than anticipated but also that they have to be chosen well because the quality does not rise linearly in any of them: choosing a too low number of, e.g., swaps can significantly harm the quality of the algorithm. A last question to be analyzed is whether it is reasonable to take the full Netflix data set with 100 million ratings or whether a random subset of it is good enough - also here we see a phase transition-like behavior.

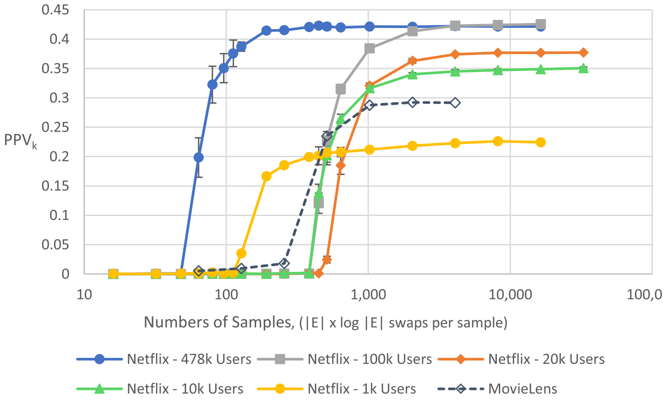

Fig. 1. Quality over number of samples. For a wide variety of datasets narrow phase transitions are present.

For the runtime of the algorithm, the most important question is how many samples need to be drawn since the co-occurrence needs to be computed for every pair of movies of interest which - in general - is in $O(n^3)$. Fig. 1 shows a very steep transition in quality in dependence of the number of samples: The full MovieLens data set requires about 1024 samples to reach a PPV_k-value of 0.286 ± 0.005. Spending more samples only marginally improves the results, e.g., for 4096 samples the PPV_k-value is 0.291 ± 0.002. For the full Netflix data set, 384 samples already result in a PPV_k-value of 0.4206 ± 0.0019, while $16,384$ samples only improve this value to 0.4217 ± 0.0012, but would take 43x more time to compute. Note, however, that this steep, phase-transition-like behavior also has the downside that a too small number of samples decreases the quality enormously: using 64 instead of 384 samples still yields a PPV_k of 0.20 ± 0.03, using only 48 samples decreases it to a useless value of 0.001 ± 0.001.

The number of swaps per sample is computationally a bit less important, but the results show again that a too low number of swaps per sample in the serial burn-in sampling scheme can strongly decrease the quality of the algorithm, even if $10,000$ samples are drawn from the random graph model. Fig. 2 shows the steep transitions when varying the number of swaps.

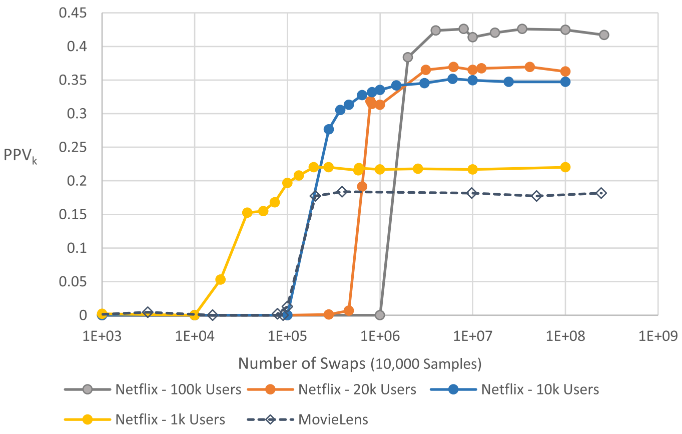

Fig. 2. Quality over number of swaps. For a wide variety of datasets narrow phase transitions are present similar to the one for samples.

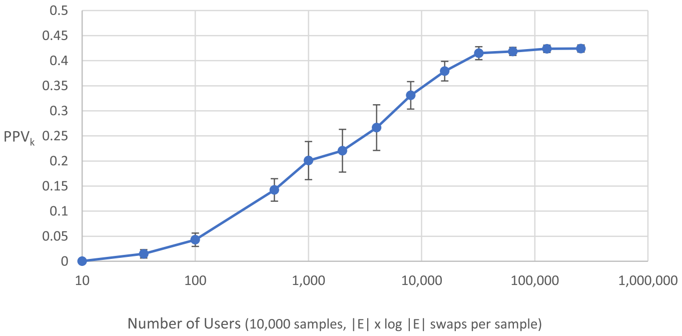

Fig. 3. Quality over the number of users in the Netflix dataset. The users are selected at random, the dataset has 478k users.

Fig. 3 shows the quality of the ranking for different sizes of subsets of users in the bipartite graph. It can be clearly seen that the maximally achievable quality depends strongly on the user set's size: with $1,000$ users, a PPV_k-value of only 0.20 ± 0.04 can be achieved while the full data set allows for a PPV_k-value of 0.424 ± 0.007. Fig. 3 shows the quality in dependence of the number of users and the number of samples from the random graph model. The data shows that taking $100,000$ users out of the $480,000$ uniformly at random (averaged over 10 of these sets) allows for the same overall quality, but only with more samples from the FDSM. It seems that, in general, a smaller set is sufficient if more samples from the random graph model are made to assess the significance of the observed co-occurrence values.

This section has shown that there are various phase transition-like changes in the resulting quality of the link assessment that need to be regarded when selecting a subset from a larger data set, the number of swaps in a serial burn-in-scheme, and finally the number of samples. While we had ground truth to evaluate this behavior in these cases, other data sets do not necessarily come with a precompiled ground truth. Thus, the next section introduces an online heuristic for large data sets that can be used to estimate the necessary number of swaps and samples to assess the significance of the observed co-occurrence in a bipartite graph.

V. Heuristics

Fig. 1 shows a sharp phase transition from a PPV_k of 0 to one of 0.41 between 48 and 192 samples; after that the PPV_k is almost constant even up to 10k samples. Since the runtime is linear in the number of samples, there is no point in generating more than 192 samples from a practical perspective. However, we also see from Fig. 1 that the transition is also data dependent. Analogously, the same observation can be made for the number of swaps. To reduce the overall runtime without decreasing the quality of the result, it is thus necessary to monitor the sampling process online and stop when the number of swaps and the number of samples is sufficient. In the following we propose two online heuristics that indicate the minimum required #samples and #swaps, without the usage of any kind of ground truth.

A. Heuristic for #swaps: Same Degree Coocc Convergence

The number of swaps needs to be high enough, otherwise the sampled graphs are not independent from the starting graph. The idea of the *swap heuristic* is to build a correlated variable θ that indicates when the Markov Chain has mixed, i.e., the built graph is independent from the starting point. For small graphs, this number can be set to $m \log m$, the number of steps such that, expectedly, every edge has been picked at least once. However, for larger graphs, this number can be prohibitively large. But larger graphs might contain a set of pairs of nodes with the same degree that start with very different co-occurrences. For example, there might be seven nodes with degree 10 and four nodes with degree 20. Thus, there are 28 pairs of nodes with the same degrees.

While in every random sample the $coocc(a, b)$ of two nodes a, b is different, we know that the average over all samples $coocc_{FDSM}(a, b)$ converges to a fixed number that only depends on the degrees of a and b. Thus, it is also the same for **all** pairs of nodes with the same degrees:

$$\begin{aligned}\forall a, b, c, d \in V_l : deg(a) = deg(c) \wedge deg(b) = deg(d)\\ \Rightarrow coocc_{FDSM}(a, b) = coocc_{FDSM}(c, d). \qquad (1)\end{aligned}$$

if the sampled graphs are independent from each other. Fig. 4 shows the $coocc_{FDSM}$ of four different movie pairs of the Netflix data, with the same degrees but different observed co-occurrences (points on the left of the x-axis). The figure then shows the average co-occurrence of $10,000$ sampled graphs in dependence of the number of swaps in the serial burn-in sampling scheme. It can be clearly seen that the average co-occurrence of all pairs converges to the same value.

Based on this insight we propose the variable $\theta(\#\text{swaps})$ to determine the optimal number of swaps as described in the following, described now: From the dataset collect all sets $D(x, y)$ of pairs of nodes that have at least N_p node pairs with the same degrees x and y. Compose G by selecting N_p pairs u.a.r. from all sets: **ToDo: [Why not use all of them? This seems to be an unjustified parameter.]**

$$\begin{aligned}D(x, y) = \{(a, b) \mid \forall a, b \in V_l : deg(a) = x, deg(b) = y\},\\ G = \{(d_1, .., d_{N_p}) \in D(x, y) \mid \forall x, y \in \mathbb{N} : |D(x, y)| \geq N_p\}.\end{aligned}$$

From this set G take N_g groups at random: $g_1, ..g_{N_g} \in G$. In each of these groups test the convergence of the average co-occurrence by computing the normalized standard deviation $\delta := s/m$ of the $coocc_{FDSM}$ in this group, where s is the standard deviation of the sample and m is its mean.

The variable θ is then defined as the mean over all deviations δ, see Fig. 4.

$$\begin{aligned}\mathcal{C}(g_i, \#\text{swaps}) &= \{coocc^{\#\text{swaps}}_{FDSM}(a, b) \mid \forall (a, b) \in g_i\},\\ \delta(g_i, \#\text{swaps}) &= \frac{std\,[\mathcal{C}(g_i, \#\text{swaps})]}{mean\,[\mathcal{C}(g_i, \#\text{swaps})]},\\ \theta(\#\text{swaps}) &= \frac{1}{N_g} \sum_{i=1}^{N_g} \delta(g_i, \#\text{swaps}).\end{aligned}$$

How good is this heuristic to find the minimal number of swaps for a high-quality result? To empirically test it, the ground truth can be used once again: Fig. 5 shows the value of θ over #swaps for $N_p = 4$ and $N_g = 24$ in blue and the PPV_k in yellow. The heuristic shows an almost inverted behavior with respect to the PPV_k, indicating a good correlation. Thus, when θ approaches 0, we assume that the quality of the resulting significance test is reliable.

Based on θ we can search now a good #swap without relying on the ground truth. Note that for small #swaps it is very efficient to evaluate θ, even for 10k samples. Starting with a low #swaps, we continuously increase it until $\theta < \theta_{min}$. A threshold of $\theta_{min} = 0.01$ seamed to be a reasonable choice in our tests, meaning an average relative error of 1% in the *co-occurrence* (*coocc*). The complete swap heuristic is shown in Algorithm 1.

```
Data: Graph $G(V_L \cup V_R, E)$ with vertices $V_L$ and $V_R$
      and edges $E$, $V_R$ being the vertices of interest,
      $N_g, N_p, \theta_{min}$, #samples;
Result: #swaps
$G_0 := G$ randomized with $|E| \ln |E|$ swaps;
Select $N_g$ groups with $N_p$ pairs of nodes with same
degrees at random from $V_R$ each;
#swaps := $\max(|V_L|, |V_R|) \ln \max(|V_L|, |V_R|)$;
$s_{high} := |E| \ln |E|$;
do
    #swaps := $\sqrt{\#\text{swaps} \cdot s_{high}}$;
    Evaluate $\theta(\#\text{swaps})$ with given #samples from $G_0$;
while $\theta(\#swaps) > \theta_{min}$;
```

Algorithm 1: Same Degree Coocc Convergence Swap Heuristic

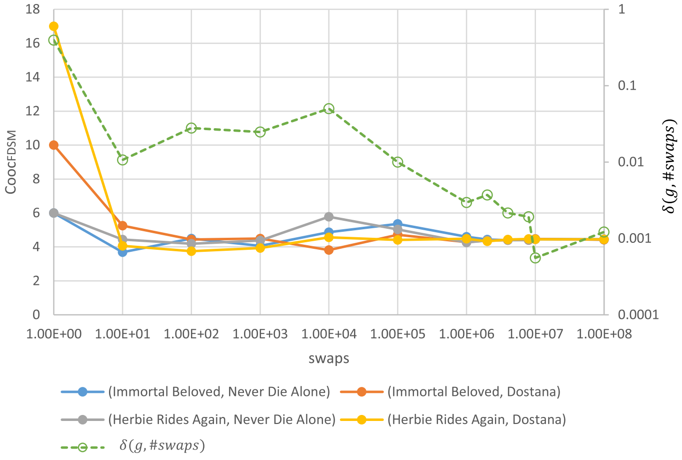

Fig. 4. Convergence of the $cooc_{FDSM}$ for the Netflix dataset with 10k users.

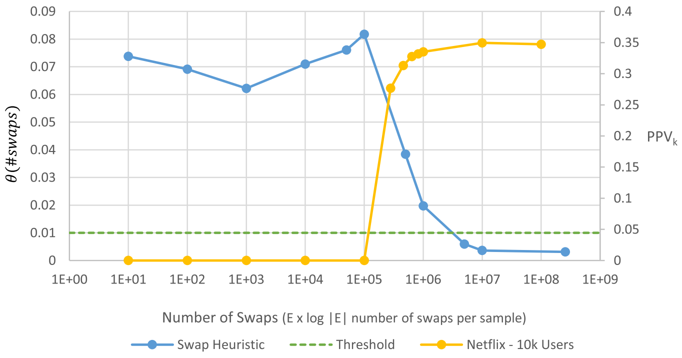

Fig. 5. Swap heuristic θ and the PPV.

B. Heuristic for #samples: Internal PPV_k

The heuristics for determining an ideal number of samples is based on the idea that the ranking of the most significant co-occurring pairs (pairs of nodes with the most significant number of common neighbors) should stabilize with the number of samples.

We thus propose to use the *internal PPV_k heuristic* which makes use of an "internal ground truth", defined as the k pairs of nodes ranked highest in the previous run, where the ranking is based on the p-value and ties are broken with respect to the z-score. Then, based on this internal ground truth GT, the PPV_k of the current result is calculated, i.e., we quantify how much the newly sampled graph(s) change the ranking of the top k' pairs and stop if that value is larger than some threshold value α. Algorithm 2 shows the steps in details.

Collecting the internal ground truth GT', while not increasing the complexity of the algorithm, is still computationally relevant. So instead of collecting it for every sample, we only collect it every $samples_{step}$ samples. The stopping quality α and the length k', in turn, should be high enough to guaranty a sufficient stability, but low enough to keep the overhead as small as possible.

Fig. 6 shows the output of the heuristic in action for the full Netflix dataset. The following configuration has been used:

$$samples_{step} = 16; \quad k' = 0.2\% \, |V_R|^2; \quad \alpha = 0.95.$$

```
Data: Graph $G(V_L \cup V_R, E)$ with vertices $V_L$ and $V_R$
      and edges $E$, $V_R$ being the vertices of interest,
      #swaps, $k'$, $samples_{step}$, $\alpha$;
Result: ranking according to $p$-value and $z$-score for all
        pairs of vertices $(u,v) \in (V_R \times V_R)$;
Calculate $coocc(u,v)\ \forall\ (u,v) \in (V_R \times V_R)$;
$G_0 := G$; $i := 0$;
do
    $GT' := k'$ pairs $(u,v)$ with the highest ranking of
    $G_i$ ;
    for $k$ := 1 to $samples_{step}$ do
        $i := i + 1$; $G_i := G_{i-1}$;
        Randomize $G_i$ with given #swap;
        Calculate $coocc_i(u,v)\ \forall\ (u,v) \in (V_R \times V_R)$;
    end
    $PPV' := PPV_{k'}$(List of $k'$ highest ranked node
    pairs containing at least one node from $GT', GT'$);
while $PPV' < \alpha$;
Calculate ranking according to Section V-A;
```

Algorithm 2: Internal PPV_k Sample Heuristic

In blue the PPV_k for the Movie Ground Truth is shown, while in dashed lines the internal PPV' is shown. And while the two curves for each data set are not perfectly aligned, the internal PPV_k converges later than the ground-truth-based PPV_k, such that it is safe to use the heuristic as a stopping point. Based on the graph we can conclude, that the internal PPV_k gives us a stable and clear indicator of when to stop the algorithm.

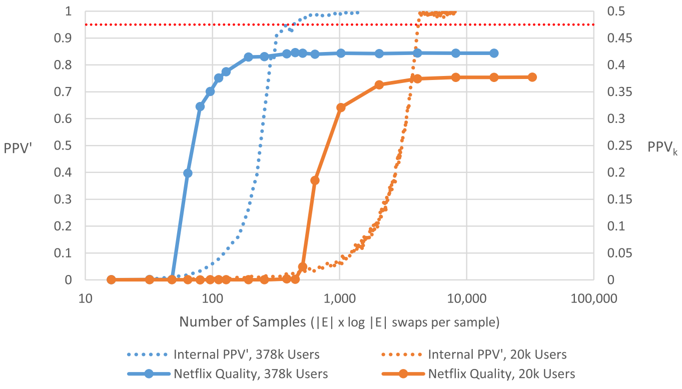

Fig. 6. Sample heuristic PPV' and PPV.

C. Results

Based on the two stopping heuristics, the runtime of the algorithm can be dramatically reduced, especially for the large Netflix data set (Table I): the swap heuristics alone seems to be only helpful for large enough graphs: the movieLens data is so small that the overhead to compute the heuristics actually increases the overall runtime. For the medium-sized Netflix subset (100 k users), the heuristics decreases the runtime by a factor of 1.5. Computing the statistical significance of all co-occurrences with a "safe" burn-in phase of $m \log m = 10^9$ swaps and the same number of swaps for each subsequent sample with a total of $10,000$ samples took 20 hours. Mind that this time was only achievable by running 128 sampling chains in parallel on a cluster, i.e., computing the algorithm on only

TABLE I. SPEED AND QUALITY FOR MOVIES GT

Data set	Samples	Swaps	Runtime	PPV
State-of-the-art				
Netflix, 487k users	10,000	10^9	20 h	0.422
Netflix, 100k users	10,000	2.6×10^8	5.5 h	0.425
MoviesLens	10,000	1.4×10^7	877 s	0.290
Swap Heuristic[2]:				
Netflix, 100k users	10,000	3.7×10^7 (**7x**)	0.2+3.4 h (1.5x)	0.424 (-0.2%)
MovieLens	10,000	1.1×10^7	370+554 s (0.95x)	0.290 ($+0\%$)
Sample Heuristic[1]:				
Netflix, 487k users	640	10^9	1.42 h (**14x**)	0.418 (-0.9%)
Movies Lens	3,456	1.4×10^7	326 s (**2.7x**)	0.291 ($+1\%$)

[1] $samples_{step} = 128$; $k' = 0.2\% \, |V_R|^2$; $\alpha = 0.95$.
[2] $N_p = 4$; $N_g = 24$; $\theta_{min} = 0.01$

one CPU would have taken approximately 3.5 months! Thus, the swap heuristic alone already decreases the computation time to an amenable runtime of 3.6 hours on the cluster. The improvement of the runtime by reducing the number of samples from $10,000$ to 640 is even larger, namely a factor of 14x, thereby reducing the quality of the link assessment by only about 1%. Even if the algorithm ran on a single CPU, the runtime would now be a bit more than a day. Similarly, also the MovieLens data profits from the swap heuristics by reducing the runtime by a factor of 2.7.

VI. CONCLUSION

While the usage of the FDSM to assess the statistical significance of structural patterns in graphs has often been proposed, its application was so far prohibitive for large-scale data. This is caused by the fact that there is no practically usable known upper bound on the mixing time of the Markov chain that allows to sample uniformly at random from the FDSM. Similarly, there is no known bound on the number of necessary samples to achieve a good quality of the *expected* values by the *observed* means in the sample.

Here we have shown that two online heuristics can help to determine a sufficient number of swaps and samples to achieve results with a very good quality with respect to the quality achievable with "safe" parameters. This was only possible by the usage of ground truth data that allows to evaluate the quality of the resulting significance evaluation.The reduction in runtime by a factor of up to 14 makes it possible to apply the proposed algorithm to the full Netflix data set and get the result within two days on a single CPU core without reducing the quality significantly. Further research will have to show whether the results can be transferred to other kind of network data like protein-protein interaction data or social network data. However, here it is much more difficult to obtain ground truth data for the evaluation.

Reducing the data to a much smaller subset is also an important achievement and we could show that a random sample of 100k users from the Netflix data set is sufficient to gain a result with almost the same quality as the much larger full data set with 480k users - however, this was only possible by using the ground truth. Further research is necessary to understand whether there is also a heuristic oblivious of any ground truth to determine the necessary size of a subset of the data to achieve a high-quality link assessment.

ACKNOWLEDGMENT

We would like to especially thank Andreas Spitz from the University of Heidelberg and Emőke-Ágnes Horvát for sharing the ground truth data set with us. Andreas Spitz also shared his Java implementation on which we built the improved version.

The simulations were executed on the high performance cluster "Elwetritsch" at the TU Kaiserslautern which is part of the "Alliance of High Performance Computing Rheinland-Pfalz" (AHRP). We kindly acknowledge the support.

REFERENCES

[1] K. A. 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 2010 International Conference on Advances in Social Networks Analysis and Mining ASONAM 2010*, 2010, pp. 200–207.

[2] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, and U. Alon, "Response to Comment on "Network motifs: Simple building blocks of complex networks " and "Superfamilies of evolved and designed networks"," *Science*, vol. 305, p. 1107d, 2004.

[3] K. A. Zweig, "Good versus optimal: Why network analytic methods need more systematic evaluation," *Central European Journal of Computer Science*, vol. 1, pp. 137–153, 2011.

[4] S. Uhlmann, H. Mannsperger, J. D. Zhang, E.-Á. Horvat, C. Schmidt, M. Küblbeck, A. Ward, U. Tschulena, K. Zweig, U. Korf, S. Wiemann, and Ö. Sahin, "Global miRNA regulation of a local protein network: Case study with the EGFR-driven cell cycle network in breast cancer," *Molecular Systems Biology*, vol. 8, p. 570, 2012.

[5] E.-Á. Horvát, J. D. Zhang, S. Uhlmann, Ö. Sahin, and K. A. Zweig, "A network-based method to assess the statistical significance of mild co-regulation effects," *PLOS ONE*, vol. 8, no. 9, p. e73413, 2013.

[6] E.-Á. Horvát and K. A. Zweig, "A fixed degree sequence model for the one-mode projection of multiplex bipartite graphs," *Social Network Analysis and Mining*, vol. 4, p. 164, 2013.

[7] A. Gionis, H. Mannila, T. Mielikäinen, and P. Tsaparas, "Assessing data mining results via swap randomization," in *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06)*, 2006.

[8] A. Gionis, et al., "Assessing data mining results via swap randomization," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 3, p. article no. 14, 2007.

[9] E.-Á. Horvát and K. A. Zweig, "One-mode projections of multiplex bipartite graphs," in *Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, 2012.

[10] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 7, pp. 1019–1031, May 2007. [Online]. Available: http://doi.wiley.com/10.1002/asi.20591

[11] R. Kannan, P. Tetali, and S. Vempala, "Simple markov-chain algorithms for generating bipartite graphs and tournaments," *Random Structures and Algorithms*, vol. 14, no. 4, pp. 293–308, 1999.

[12] M. Jerrum, A. Sinclair, and E. Vigoda, "A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries," *Journal of the ACM (JACM)*, vol. 51, no. 4, pp. 671–697, 2004.

[13] R. A. Brualdi, "Matrices of zeros and ones with fixed row and column sum vectors," *Linear Algebra Applied*, vol. 33, pp. 159–231, 1980.