Efficiently covering complex networks with cliques of similar vertices

Michael Behrisch [1,*]

Institut für Informatik, Humboldt-Universität zu Berlin, 10099 Berlin, Germany

Anusch Taraz [2]

Zentrum Mathematik, Technische Universität München, 80290 München, Germany

Abstract

We describe a polynomial time algorithm for covering graphs with cliques, prove its asymptotic optimality in a random intersection graph model and present experimental results on complex real–world networks.

1 Introduction

The construction of stochastical models for complex real–world networks of huge dimensions has attracted an enormous amount of attention during the last five years. These efforts are motivated by several aspects, namely the prediction of network structure as well as the design, benchmarking and theoretical verification of algorithms.

As *graphs* are the canonical model for networks, *random graphs* seem to be appropriate candidates for the stochastic models. The classical random graph model was introduced by Erdős and Rényi in the early 1960s. It is denoted by $G_{n,p}$ and considers a fixed set of n vertices and edges that exist with a certain probability $p = p(n)$, independently from each other. However, it lacks many

* Corresponding author.
Email addresses: behrisch@informatik.hu-berlin.de (Michael Behrisch), taraz@ma.tum.de (Anusch Taraz).

[1] supported by the DFG research center "Mathematics for key technologies" (FZT 86) in Berlin.
[2] supported in part by the DFG research center "Mathematics for key technologies" (FZT 86) in Berlin.

of the commonly observed properties of real–world networks (e.g. scale free degree distribution and clustering). One of the underlying reasons that are responsible for this mismatch is precisely the independence of the edges, in other words the missing transitivity. In a real–world network, relations between vertices x and y on the one hand and between vertices y and z on the other hand suggest a connection of some sort between vertices x and z.

To take better care of this fact, we will investigate random intersection graphs. An *intersection graph* is a graph $G = (V, E)$ together with a so–called *universal feature set* W. Every vertex $x \in V$ has an assigned *feature set* $W_x \subseteq W$, and the characteristic property of an intersection graph is that two vertices $x, y \in V$ are connected by an edge in E if and only if their feature sets have non-empty intersection:

$$\{x, y\} \in E \Leftrightarrow W_x \cap W_y \neq \emptyset.$$

We call the elements of W *features*. If the feature $w \in W$ is contained in W_x and W_y and thus forces the edge $\{x, y\}$, we say that $\{x, y\}$ is *induced* by w. Furthermore the set of vertices V_w holding a specified feature w (which obviously forms a clique) is called a *feature clique*. Trivially

$$v \in V_w \Leftrightarrow w \in W_v,$$

in which case we say that v and w *see* each other.

Examples for intersection graphs are interval graphs (see e.g. [1]), where the feature sets consist of intervals on the real line. In this paper however we will only consider finite sets.

A *random intersection graph* on n vertices with a universal feature set of size m is a probability model where each vertex chooses each feature independently with probability p. A sample of this probability space is denoted by $G_{n,m,p}$.

A few simple observations. Obviously $G_{n,m,p}$ does exhibit some kind of transitivity: if the edges $\{x, y\}$ and $\{y, z\}$ are induced by the same feature w, then this will also induce the edge $\{y, z\}$. The smaller m is, the 'simpler' will be $G_{n,m,p}$, because relatively few cliques will dominate its structure. In the following we will consider the case $m := n^\alpha$. It was shown in [2] that for $\alpha > 6$ the random intersection graph $G_{n,m,p}$ behaves in many ways like the classical random graph $G_{n,p'}$ with $p' = 1-(1-p^2)^m$. We will focus in this paper mainly on the case where $0 < \alpha < 1$.

It is sometimes convenient to view the random intersection graph as a random bipartite graph with bipartition (V, W) and random edges between the V and W occurring independently with probability p. A sample from this space will be denoted by $B_{n,m,p}$. Given the bipartite graph, say B, the intersection graph

is obtained as $G = B^2[V]$, where we write B^2 for the so–called square of B (where two vertices are connected if their distance is at most 2 in B). B is called a *generator* of G.

The model of a random intersection graph $G_{n,m,p}$ has been studied with respect to subgraph appearance by Karoński, Scheinerman and Singer-Cohen in [3] and with respect to equivalence to $G_{n,p}$ by Fill, Scheinerman, Singer-Cohen in [2] (see also [4]). Stark has investigated the vertex degree distribution in [5]. The evolution of the largest component, for growing p and fixed α has been studied by the first author in [6]. Extensions to the model have been proposed by Godehardt and Jaworski in [7], who modify the distribution of the sizes of the feature cliques. The practical relevance of the model has been discussed by Newman, Strogatz and Watts in [8] and by Guillaume and Latapy in [9].

The main aim of this paper is to develop and analyze simple algorithms which, given an intersection graph, quickly reproduce the underlying feature cliques. As the features of a network are likely to reflect important properties of the data, they represent important meta-information that will help in clustering, storing and searching it efficiently. An immediate example for such feature cliques are *common topics* in a network containing *documents* as vertices and edges between *similar documents*.

Since every graph can be seen as an intersection graph with the universal feature set being large enough, we want to (re)produce a universal feature set that is as small as possible. This is equivalent to the NP-hard problem of constructing an (edge) clique cover with a minimum number of cliques for the graph [10], and hence we cannot expect to find an efficient algorithm which always finds an optimal solution. Instead, we present a simple greedy heuristic that constructs the generator of a given graph. Our main contribution is to prove that this algorithm performs *a.a.s.* optimally (this means with probability tending to one as n tends to infinity), when the input graph is chosen at random from our model $G_{n,m,p}$ for certain ranges of p. More precisely, we will prove the following two theorems.

Theorem 1 *Let a positive constant $\alpha < 1$, n, $m := n^\alpha$ and $\frac{\ln^2 n}{n} \leq p = O(\frac{1}{m})$ be given and let $G := G_{n,m,p} = (V, E)$ be a random intersection graph. Then there exists an algorithm which a.a.s. finds in $O(n^3)$ a bipartite graph $B = (V \cup W, A)$ with $|W| \leq m$ and $B^2[V] = G$ (a generator of G).*

Theorem 2 *Let a positive constant $\alpha < 1$, n, $m := n^\alpha$ and $\frac{\ln^2 n}{n} \leq p < \min\{\frac{1}{5}m^{-\frac{2}{3}}, \frac{n}{8m^2}\}$ be given and let $G := G_{n,m,p} = (V, E)$ be a random intersection graph. Then there exists an algorithm which a.a.s. finds in polynomial time a bipartite graph $B = (V \cup W, A)$ with $|W| \leq m$ and $B^2[V] = G$ (a generator of G).*

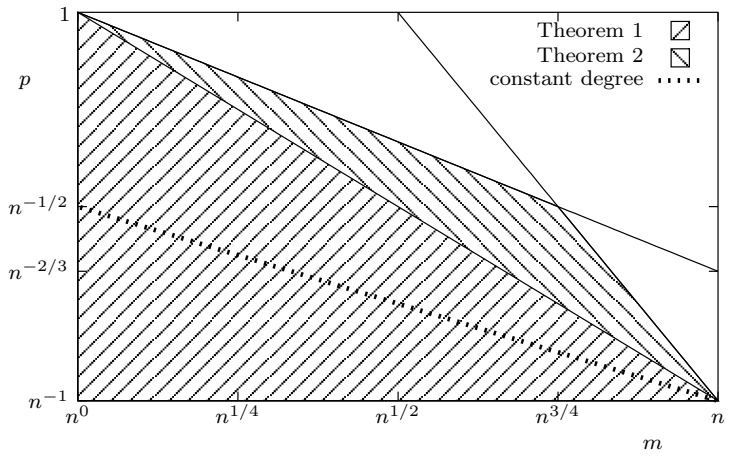

Fig. 1. Ranges for p and m for which we prove the a.a.s. optimality of Algorithm 1

Notice that Theorem 2 covers a greater range of p at the expense of a larger (but still polynomial) running time of the algorithm. Observe that in particular graphs with constant expected degree (which seems appropriate for many real–world networks) are already covered by Theorem 1 and can thus be analyzed very efficiently. Figure 1 illustrates the range of m and p for which our theorems hold.

Following Guillaume and Latapy [9], who compared real complex networks with random intersection graphs, we ran our algorithm on the same or similar real–world networks to obtain a clique cover. The simulation results show that even very large graphs can be covered quite well with a reasonable number of cliques and a good running time. More importantly, these experiments suggest values for m (and thus, via the edge density, also values for p), and enabled us to compare the degree distribution in individual real-world networks with those in the random intersection graph $G_{n,m,p}$ with the "correct" parameters m and p.

This paper is organized as follows. Section 2 contains the algorithm that gives rise to the theorems. Section 3 presents some basic lemmas concerning the random intersection graph model. In Section 4 we prove Theorem 1 which is just a warmup for the proof of Theorem 2 in Section 5. We close with some experimental results and a comparison of some properties in real networks and our random graph model.

2 The algorithm

The following algorithm finds cliques in a graph by testing the common neighborhood of vertex subsets of fixed size k for completeness. From the cliques found in this way it takes the largest ones in order to cover the graph.

We shall use the following (slightly non–standard) notation: For the set $A \cup \{x\}$ we write $A + x$. Denote by $\Gamma(v)$ the set of vertices having edges to v and by $N(v) := \Gamma(v) + v$ the same set including v itself. For a vertex set U we denote by $Z(U)$ the common neighborhood of the vertices in U ($Z(U) := \bigcap_{i=1}^{k} N(v_i)$).

Algorithm 1
Input: *Graph G on n vertices, $k \in \mathbb{N}$*
Output: *(partial) edge clique cover $\mathcal{M}$ of G*
FEATUREFIND*(G, k)*
(1) $\mathcal{L} := \emptyset$;
(2) ***foreach*** $U_k = \{v_1, \ldots, v_k\} \subseteq V$
(3) $Z = Z(U_k) := \bigcap_{i=1}^{k} N(v_i)$
(4) ***if*** $G[Z]$ *complete*
(5) $\mathcal{L} := \mathcal{L} + Z$;
(6) $Y := \emptyset$;
(7) ***foreach*** $Z \in \mathcal{L}$ *in decreasing cardinality* $|Z|$
(8) ***if*** $E(G[Z]) \not\subseteq Y$
(9) $Y := Y \cup E(G[Z])$;
(10) $\mathcal{M} := \mathcal{M} + Z$;

We will use this algorithm with $k = 1$ to prove Theorem 1 and with larger k to prove Theorem 2. The set $\mathcal{M}$ found by the algorithm contains the vertex sets seen by the individual features and can thus be considered as a subset of the feature set W of a possible generator of G.

The running time of the algorithm is clearly dominated by checking the clique property for the neighborhood of all k-subsets of V which leads to a total of $O\left(\binom{n}{k} n^2\right)$. The following proposition gives rise to an algorithm which needs much less time in practice.

Proposition 3 *Let $G = (V, E)$ be a graph and let $U \subseteq V$ be such that $C := Z(U) = \bigcap_{u \in U} N(u)$ is a clique in G. Furthermore let U' be an arbitrary subset of C. If $Z(U')$ is a clique then $Z(U') = C$.*

PROOF. Since C is a clique it is immediate that for every subset $U' \subseteq C$ all vertices of C are adjacent to all vertices of U', hence $C \subseteq Z(U')$. Now assume that $Z(U')$ is a clique and that there is a vertex v in $Z(U')$ which is not in C. Since $C \subseteq Z(U')$ all vertices in C (and especially in U) are adjacent to v but this means $v \in Z(U) = C$ which contradicts the assumption that $v \notin C$. Thus v cannot exist and the statement is proven. $\square$

This proposition implies that every set U_k which is a subset of a clique that has been found in an earlier stage of the algorithm does not have to be checked

anymore, which in practice reduces the number of sets to be checked dramatically.

Furthermore note that for $k = 1$ (and in fact even for $k = 2$) sorting the cliques (starting at line 7) and taking only the largest ones is not necessary because the way in which we find them already ensures that a clique in $\mathcal{L}$ contains at least one vertex (resp. edge for $k = 2$) which is in no other clique in $\mathcal{L}$. Thus $\mathcal{M}$ and $\mathcal{L}$ are equal.

3 Auxiliary Lemmas

The following estimates are used without proof:

$$\binom{a}{b} \leq \left(\frac{\mathrm{e}a}{b}\right)^b \tag{1}$$

$$\binom{a}{b} \leq a^b \tag{2}$$

$$1 - ab \leq (1-a)^b \leq 1 - \frac{ab}{2} \qquad \text{for } 0 \leq a \leq 1, ab < 1 \tag{3}$$

$$\mathrm{e}^{-2a} \leq 1 - a \leq \mathrm{e}^{-a} \qquad \text{for } 0 \leq a \leq \frac{1}{2} \tag{4}$$

Let X be a non-negative random variable with expectation $\mu = \mathbb{E}[X]$. As a special case of Markov's inequality the first moment method states that

$$\mathbb{P}[X \geq 1] \leq \mu. \tag{5}$$

If X is binomially distributed random variable (n trials, each with probability p), then $\mu = np$ and we shall use the following variants of Chernoff's inequality (see Section 2 in [11]):

$$\mathbb{P}[X \geq \mu + t] \leq \exp\left(-\frac{t^2}{2(\mu + t/3)}\right) \qquad \text{for } t \geq 0, \tag{6}$$

$$\mathbb{P}[X \leq \mu - t] \leq \exp\left(-\frac{t^2}{2\mu}\right) \qquad \text{for } t \geq 0, \tag{7}$$

$$\mathbb{P}[X \geq t] \leq \exp(-t) \qquad \text{for } t \geq 7\mu. \tag{8}$$

Let $G_{n,m,p}$ be a random intersection graph. We first show that the probability that there is a feature clique which deviates much from its expected size is exponentially small.

Lemma 4 *Let $X_w := |V_w|$ be the random variable counting the number of vertices of a fixed feature w in a random intersection graph $G_{n,m,p}$ with $m := n^\alpha$*

and $\alpha < 1$. *Then*

$$\mathbb{P}\left[\exists w \in W : |X_w - pn| > \frac{pn}{2}\right] \leq 2m\mathrm{e}^{-\frac{pn}{10}}.$$

PROOF. The number of vertices seen by a feature is a binomially distributed random variable with expected value np. For a fixed feature w we have

$$\mathbb{P}\left[X_w > pn + \frac{pn}{2}\right] \overset{(6)}{\leq} \exp\left(-\frac{(pn)^2}{8(pn + \frac{pn}{6})}\right) \leq \mathrm{e}^{-\frac{pn}{10}}$$

$$\mathbb{P}\left[X_w < pn - \frac{pn}{2}\right] \overset{(7)}{\leq} \exp\left(-\frac{(pn)^2}{8pn}\right) \leq \mathrm{e}^{-\frac{pn}{10}}.$$

Using linearity of expectation (summing over all features w) and the first moment method we obtain that

$$\mathbb{P}\left[\exists w \in W : |X_w - pn| > (pn)^{\frac{3}{4}}\right] \leq 2m\mathrm{e}^{-\frac{pn}{10}}.$$

□

Similar results hold for the size of the feature sets.

Lemma 5 *Let* $X_v := |W_v|$ *be the random variable counting the number of features for a fixed vertex* v *in a random intersection graph* $G_{n,m,p}$ *with* $m := n^\alpha$ *and* $\alpha < 1$. *Then*

$$\mathbb{P}\left[\exists v \in V : X_v > 2pm\right] \leq n\mathrm{e}^{-\frac{3pm}{8}},$$

and for $pm \leq 3\ln n$

$$\mathbb{P}\left[\exists v \in V : X_v > 21\ln n\right] \leq \frac{1}{n^{20}}.$$

PROOF. Very similarly to the previous lemma, we have for a fixed vertex v

$$\mathbb{P}\left[X_v > pm + pm\right] \overset{(6)}{\leq} \exp\left(-\frac{(pm)^2}{2(pm + pm/3)}\right) = \mathrm{e}^{-\frac{3pm}{8}}$$

and for $pm \leq 3\ln n$

$$\mathbb{P}\left[X_v > 21\ln n\right] \overset{(8)}{\leq} \exp(-21\ln n) = \frac{1}{n^{21}}.$$

Again summing over all vertices v yields the statement of the lemma. □

Denote by B the event that none of the events in Lemmas 4 and 5 occur. In other words, for no $w \in W : |X_w - pn| > \frac{pn}{2}$ and for no $v \in V : X_v > 2pm$ or $X_v > 21 \ln n$. The above lemmas show that (under certain conditions on n, m and p) we have $\mathbb{P}\left[\bar{B}\right] \to 0$. In the following we will often observe that these conditions are indeed satisfied, and then attempt to compute the probability for some other event A. As

$$\mathbb{P}[A] = \mathbb{P}[A|B]\mathbb{P}[B] + \mathbb{P}\left[A|\bar{B}\right]\mathbb{P}\left[\bar{B}\right] \leq \mathbb{P}[A|B] + \mathbb{P}\left[\bar{B}\right],$$

we can then restrict our attention to proving that $\mathbb{P}[A|B] \to 0$.

4 The case $k = 1$

We first show that almost surely every feature clique contains a vertex with only one feature.

Lemma 6 *Let $G_{n,m,p}$ with $m := n^\alpha$, $\alpha < 1$ and $\frac{\ln^2 n}{n} \leq p = O(\frac{1}{m})$ be a random intersection graph. Then a.a.s. every feature clique V_w contains a vertex for which w is the only feature:*

$$\forall w \in W \; \exists v \in V_w : \; W_v = \{w\}.$$

PROOF. For $p \geq \frac{\ln^2 n}{n}$ we know from Lemma 4 that we can condition on the event that there is a.a.s. no feature clique with less than $\frac{pn}{2}$ vertices. Now fix a single feature w, let it choose its clique V_w and determine the probability that all the vertices inside V_w choose another feature. Summing over all features we can then bound the probability for the existence of such a w by

$$\begin{aligned}
\mathbb{P}[\exists w \in W, \forall v \in V_w : |W_v| > 1] &\leq m\left(1 - (1-p)^{m-1}\right)^{\frac{pn}{2}} \\
&\overset{(4)}{\leq} m\left(1 - \mathrm{e}^{-2pm}\right)^{\frac{pn}{2}} \\
&\leq m\left(1 - \mathrm{e}^{-O(1)}\right)^{\frac{pn}{2}} \\
&\leq m\left(1 - \mathrm{e}^{-O(1)}\right)^{\frac{\ln^2 n}{2}}.
\end{aligned}$$

This tends to 0 because for n large enough $\ln(1 - \mathrm{e}^{-O(1)}) \ln n < -\alpha$. $\square$

Theorem 1 now follows immediately from this lemma because Algorithm 1 only needs to "find" the vertex from Lemma 6 which it will surely achieve running with $k = 1$.

Proof of Theorem 1 We run Algorithm 1 with $k = 1$ and claim that a.a.s. the produced list $\mathcal{L}$ will contain exactly the feature cliques. By Lemma 6 we can assume that every feature clique V_w contains a vertex u_w for which w is the only feature ($W_u = \{w\}$). Observe that for such a vertex u_w the neighborhood $N(u_w)$ is a feature clique. This already implies that all feature cliques will be contained in $\mathcal{L}$.

Now assume that there is a vertex v with more than one feature (e.g. $x, y \in W_v$). Since u_x and u_y must lie in $N(v)$ (because v shares one feature with each of them) and since there is no edge between u_x and u_y (they have only one feature) $N(v)$ cannot be a clique. Thus if $N(v)$ is a clique, then this implies that $v = u_w$ for some feature w, and therefore $\mathcal{L}$ contains exactly the feature cliques.

The running time is bounded from above by the time needed to check the clique property for at most n sets which can surely be done in $O(n^3)$. □

Theorem 1 covers already a significant portion of interesting intersection graphs, in particular graphs with expected constant degrees (linear number of edges) and with a giant component. Both properties occur when $p = c/\sqrt{mn}$ (see [6] for details).

5 The case $k > 1$

The proof of Theorem 2 needs some more lemmas because the a.a.s. existence of a vertex with only one feature cannot be guaranteed for larger p. We will use two other asymptotic properties of the feature cliques instead. First we prove that feature cliques are maximal with respect to inclusion (Lemma 7) and from this deduce that in fact there are no larger cliques in the graph (Lemma 9). Together with the a.a.s. existence of at least one set U_k whose common neighborhood $Z(U_k)$ is complete (Lemma 8) this will prove the theorem.

Lemma 7 *Consider $m := n^\alpha$, $\alpha < 1$, a positive constant k and a random intersection graph $G_{n,m,p}$ with $\frac{k}{m} \leq p < \frac{1}{\sqrt{m}\ln n}$. Then a.a.s. every feature clique is inclusion maximal:*

$$\forall w \in W \; \forall v \in V : \; V_w \not\subseteq \Gamma(v).$$

PROOF. First observe that the statement of the formula is trivial for $v \in V_w$ since no vertex can be part of its own neighborhood. Now assume that we have the bounds on the sizes of the feature cliques and sets from Lemma 4 and Lemma 5. Suppose that for some vertex w there exists a vertex $v \notin V_w$

with $V_w \subseteq \Gamma(v)$. We will show that the probability of this event vanishes. First consider the case where $pm > 3 \ln n$:

$$\begin{aligned}
\mathbb{P}\left[\exists w \in W, v \in V : V_w \subseteq \Gamma(v)\right] &\leq mn \sum_{i=1}^{2pm} \binom{m}{i} p^i \left(1-(1-p)^i\right)^{\frac{pn}{2}} \\
&\overset{(1)(3)}{\leq} mn \sum_{i=1}^{2pm} \left(\frac{\mathrm{e}mp}{i}\right)^i (pi)^{\frac{pn}{2}} \\
&\leq mn \sum_{i=1}^{2pm} \left(\frac{\mathrm{e}mp}{i}\right)^{\frac{pn}{2}} (pi)^{\frac{pn}{2}} \qquad \text{with } i < \mathrm{e}mp < \frac{pn}{2} \\
&\leq mn 2pm (\mathrm{e}mp^2)^{\frac{pn}{2}} \\
&\leq mn 2pm \left(\frac{\mathrm{e}}{\ln^2 n}\right)^{\frac{pn}{2}},
\end{aligned}$$

which tends to 0 because $\frac{\mathrm{e}}{\ln^2 n} \to 0$ and $pn \geq n^{1-\alpha}$.

Now for the case where $pm \leq 3 \ln n$:

$$\begin{aligned}
\mathbb{P}\left[\exists w \in W, v \in V : V_w \subseteq \Gamma(v)\right] &\leq mn \sum_{i=1}^{21 \ln n} \binom{m}{i} p^i \left(1-(1-p)^i\right)^{\frac{pn}{2}} \\
&\overset{(2)(3)}{\leq} mn \sum_{i=1}^{21 \ln n} (mp)^i (pi)^{\frac{pn}{2}} \\
&\leq mn \sum_{i=1}^{21 \ln n} (p^2 mi)^{\frac{pn}{2}} \\
&\leq 21 mn \ln n (21 p^2 m \ln n)^{\frac{pn}{2}} \\
&\leq 21 mn \ln n \left(\frac{21}{\ln n}\right)^{\frac{pn}{2}},
\end{aligned}$$

which tends to 0 because $\frac{21}{\ln n} \to 0$ and $pn \geq n^{1-\alpha}$. $\square$

Now we prove that we can indeed find the feature cliques with our algorithm.

Lemma 8 *Let $\varepsilon > 0$ be fixed and consider $m := n^\alpha$, $\alpha < 1$, an integer $k > \frac{\alpha+1}{2\alpha\varepsilon}$ and a random intersection graph $G_{n,m,p}$ with $\frac{k}{m} \leq p < m^{-\frac{1}{2}-\varepsilon}$. Then a.a.s. every feature clique has a subset U_k of size k such that $V_w = Z(U_k)$ (with Z being defined in the algorithm).*

PROOF. Fix a feature w and let U_k be a fixed k-clique with $U_k \subseteq V_w$ (remember that all subsets of V_w are cliques). Furthermore let $v \in V_w$ be an arbitrary vertex. As V_w is a clique, $U_k \subseteq N(v)$ which is equivalent to $v \in \bigcap_{i=1}^k N(u_i) = Z(U_k)$. Thus $v \in V_w$ and, because v was chosen arbitrarily,

$V_w \subseteq Z(U_k)$. If $Z(U_k)$ is complete we know from Lemma 7 that $Z(U_k) = V_w$ and we are done.

So assume the opposite, e.g. there are $x, y \in Z(U_k)$ which are not adjacent. Since V_w is a clique, x or y has to be outside of V_w. Let us assume it is x, then the event of $Z(U_k)$ being not complete implies the event that there exists an $x \in Z(U_k) \setminus V_w$. This means there is an x that is in the neighborhood of all vertices in U_k but does not see feature w.

We bound the probability for this event by summing over all possible sets of (at most k) features which connect x and U_k.

$$\begin{aligned}
\mathbb{P}\left[\exists x \in V \setminus V_w \forall u \in U_k : x \in \Gamma(u)\right] &\leq n \sum_{i=1}^{k} \binom{m}{i} p^i (1-(1-p)^i)^k \\
&\overset{(1)(3)}{\leq} n \sum_{i=1}^{k} \left(\frac{epm}{i}\right)^i (pi)^k \\
&\leq n \sum_{i=1}^{k} (ep^2 m)^k \qquad \text{with } i \leq k \leq pm \\
&= nk(ep^2 m)^k.
\end{aligned}$$

If this tends to 0, a subset U_k will a.a.s. have $Z(U_k) = V_w$ for our fixed w. In order to have this for all w, we need

$$mnk(ep^2 m)^k \to 0,$$

which happens indeed for $k > \frac{\alpha+1}{2\alpha\varepsilon}$. $\square$

Finally we state that the sorting step at the end of the algorithm will indeed list the feature cliques first. In order to do so, we prove that a.a.s. all large cliques are feature cliques.

Lemma 9 *Consider a random intersection graph $G_{n,m,p}$ with $m := n^\alpha$ and $\frac{k}{m} \leq p < \min\{\frac{1}{5}m^{-\frac{2}{3}}, \frac{n}{8m^2}\}$ for some constant k. Then a.a.s. every clique of size at least $\frac{pn}{2}$ is a feature clique:*

$$\forall S \subseteq V \text{ with } |S| > \frac{pn}{2} \text{ and } G[S] \text{ is complete } : \exists w \in W \text{ such that } S \subseteq V_w.$$

PROOF. Assume that the statement of the lemma is wrong. Thus there exists a clique S of size $\frac{pn}{2} + 1$ which is no feature clique. Let $s \in S$ be an arbitrary vertex in S. Again, we first consider the case where $pm > 3 \ln n$. From Lemma 5 we know that a.a.s. no vertex in V has more than $2pm$ features, so this applies to s, too. But since s has $\frac{pn}{2}$ neighbors, there has to exist a subset

$X \subseteq N(s)$ of size $\frac{pn}{4pm} = \frac{n}{4m}$ which shares a common feature w (by the pigeon hole principle). Furthermore there has to exist a vertex $v \in S$ with $v \notin V_w$, otherwise S would be inside a feature clique. We now bound the probability of the existence of such an X and v with $X \subseteq \Gamma(v)$ (remember that S is a clique). Here we use that by Lemma 4 the size of V_w is a.a.s. at most $2pn$ and by Lemma 5 $|W_v| \leq 2pm$.

$$
\begin{aligned}
&\mathbb{P}\left[\exists w \in W, v \in V, X \subseteq V_w : |X| = \frac{n}{4m} \wedge X \subseteq \Gamma(v)\right] \\
&\quad\leq mn\binom{2pn}{|X|}\sum_{i=1}^{2pm}\binom{m}{i}p^i\left(1-(1-p)^i\right)^{|X|} \\
&\quad\overset{(1)}{\leq} mn\binom{2pn}{\frac{n}{4m}}\sum_{i=1}^{2pm}\left(\frac{emp}{i}\right)^i\left(1-(1-p)^i\right)^{\frac{n}{4m}} \\
&\quad\overset{(1)(3)}{\leq} mn(8epm)^{\frac{n}{4m}}\sum_{i=1}^{2pm}\left(\frac{emp}{i}\right)^i(pi)^{\frac{n}{4m}} \\
&\quad\leq mn(8epm)^{\frac{n}{4m}}\sum_{i=1}^{2pm}\left(\frac{emp}{i}\right)^{\frac{n}{4m}}(pi)^{\frac{n}{4m}} \qquad \text{with } i < 2pm < \frac{n}{4m} \\
&\quad= mn(8epm)^{\frac{n}{4m}}2pm(ep^2m)^{\frac{n}{4m}} \\
&\quad= 2pm^2n(8e^2p^3m^2)^{\frac{n}{4m}} \\
&\quad\leq 2pm^2n\left(\frac{72}{125}\right)^{\frac{n}{4m}},
\end{aligned}
$$

which tends to 0.

For the case where $pm \leq 3\ln n$ Lemma 5 only gives a bound of $21\ln n$ on the size of the feature set. With the same considerations as above this leads to a set X of size $\frac{pn}{42\ln n}$ and hence:

$$
\begin{aligned}
&\mathbb{P}\left[\exists w \in W, v \in V, X \subseteq V_w : |X| = \frac{pn}{42\ln n} \wedge X \subseteq \Gamma(v)\right] \\
&\quad\leq mn\binom{2pn}{\frac{pn}{42\ln n}}\sum_{i=1}^{21\ln n}\left(\frac{emp}{i}\right)^i\left(1-(1-p)^i\right)^{\frac{pn}{42\ln n}} \\
&\quad\overset{(1)(3)}{\leq} mn(84e\ln n)^{\frac{pn}{42\ln n}}\sum_{i=1}^{21\ln n}(mp)^i(pi)^{\frac{pn}{42\ln n}} \\
&\quad\leq mn(84e\ln n)^{\frac{pn}{42\ln n}}\sum_{i=1}^{21\ln n}(p^2mi)^{\frac{pn}{42\ln n}} \\
&\quad\leq mn(84e\ln n)^{\frac{pn}{42\ln n}}21\ln n(21p^2m\ln n)^{\frac{pn}{42\ln n}} \\
&\quad= 21mn\ln n(1764ep^2m\ln^2 n)^{\frac{pn}{42\ln n}} \\
&\quad\leq 21mn\ln n(80em^{-1/3}\ln^2 n)^{\frac{pn}{42\ln n}},
\end{aligned}
$$

which tends to 0. □

The proof of Theorem 2 now merely requires collecting the statements of the lemmas.

Proof of Theorem 2 We make a case distinction over p. For $\frac{\ln^2 n}{n} \leq p = O(\frac{1}{m})$ we already know from Theorem 1 that the statement is true.

Now let $k := 6/\alpha$ and consider $\frac{k}{m} < p < \frac{1}{5}m^{-\frac{2}{3}}$. Set $\varepsilon = 1/6$ and apply Lemma 8: a.a.s. for each feature $w \in W$ there exists a set $U_k(w)$ with $Z(U_k(w)) = V_w$. Hence all feature cliques will be listed in $\mathcal{L}$ after running the algorithm with k chosen as above.

Since we know from Lemma 4 that there is a.a.s. no feature clique with less than $\frac{pn}{2}$ vertices and from Lemma 9 that all cliques with more than $\frac{pn}{2}$ vertices are feature cliques we can conclude that sorting the list of cliques by their size and taking the elements until the graph is covered will a.a.s. succeed in reconstructing a bipartite graph which generates our input graph as intersection graph.

Again the running time of our algorithm is bounded by the time needed to check the clique property for the joint neighborhood of all subsets of size k, and thus $O\left(\binom{n}{k}n^2\right)$. □

6 Simulation

We tested our algorithm with seven real–world networks from different application areas. The "Mercator" graph is a graph of the internet at router level taken from [12]. The next four graphs are the same as in [9]. "Internet" describes part of the internet computer network, "Web" is the link graph of a complex website, "Authors" denotes a coauthoring graph and "Proteins" is an interaction graph of proteins. For details see [9] and [13]. Moreover "DIP" stands for "Dictionary of Interfaces in Proteins" and is a similarity graph of protein parts (vertices are protein interfaces that are adjacent if they are similar) studied in [14]. "Drugs" is the result of a search for "relatives" of 13 substances in a database of 2000 drugs where we put an edge for each pair of drugs which are relatives to the same test substance. The importance of this search is described in detail in [15].

To test the algorithm we started it on each graph with different values of k. In two cases we knew in advance the number of features that generated our graph (namely for "Authors" where the publications are the features, and for "Drugs" where the test substances are the features) which should be an upper bound of the number of cliques the algorithm needs to cover the graph.

	Mercator	Internet	Web	Authors	Proteins	DIP	Drugs
n	284805	75885	325729	16400	2113	5119	2000
$\|E\|$	449246	357317	1090108	29552	2203	14434	163969
m_{in}	–	–	–	19885	–	–	13
$\|\mathcal{M}\|$	366135	246725	425058	11710	1937	3307	11
α	1.0200	1.1049	1.0210	0.9653	0.9886	0.9488	0.3713
coverage	96.1%	93.0%	90.5%	99.6%	100.0%	80.4%	99.9%
$p \cdot 10^6$	5.500	22.428	6.953	137.00	714.11	577.39	88035.9

Table 1
Statistics on the performance of the algorithm on seven real–world networks

Table 1 gives statistics on the algorithm performance on each graph measured in the number of cliques ($|\mathcal{M}|$) that were needed to cover almost the whole graph (the "coverage" fraction of the edges is given, too) and the values of p and α resulting from this coverage.

The algorithm was run in all cases with $k = 2$, which produced a considerably better coverage than $k = 1$ while larger $k > 2$ gave only small improvements. The only exception was the "DIP" graph for which we obtained a coverage of 89.5% with 3232 features for $k = 3$.

As one can see, it is possible to cover a large portion of the graph with a number of cliques that is considerably smaller than the number of edges and also smaller than the number of cliques needed by the algorithm in [9] (which covered the whole graph).

In order to give further evidence for the adequacy of our model we compared the degree distribution for small degrees of the original real–world networks and our theoretical prediction based on the degree distribution of random intersection graphs calculated in [6]. The results are shown in Figure 2.

Especially for smaller graphs and smaller degrees the approximation is quite good. Of course it is not quite as good as that in [9], but this is due to the fact that there the whole degree distribution was used as an input, whereas we only have the two parameters p and m to adjust the model.

For the "Drugs" database the theoretical predicted degrees are smaller than the experimental ones. This is due to the so–called "bipartite clustering" (as described in [9]) which in our case means that the features are not completely independent but somewhat transitive, as there are "similar" features. This results in a larger overlap between some feature cliques than is theoretically predicted and thus leads to larger degrees of the vertices involved.

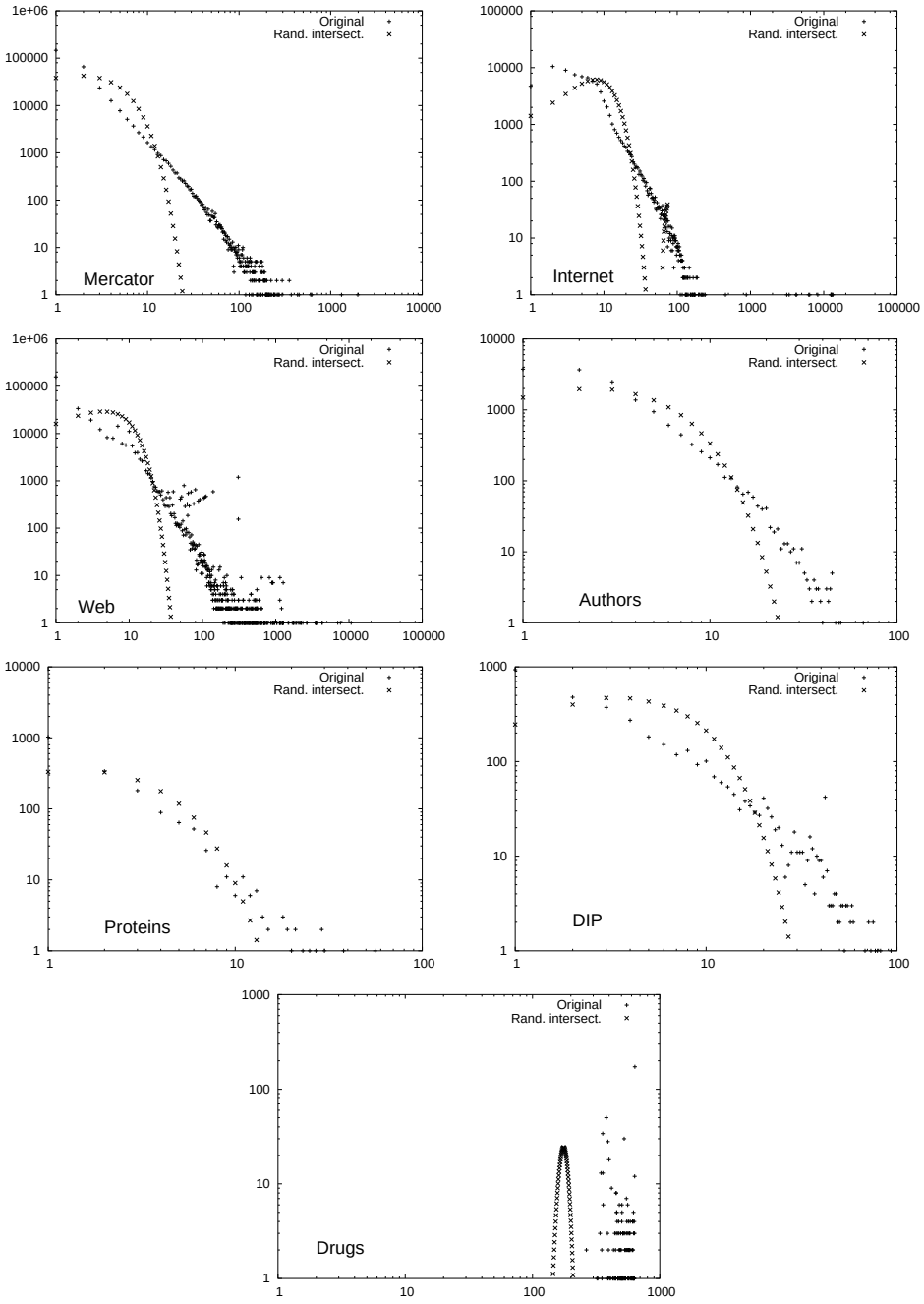

Fig. 2. Degree distributions for real–world networks: experimental results and theoretical predictions

7 Conclusion and acknowledgment

Our analysis yields a rigorous proof for the asymptotic optimality of our simple greedy algorithm in the random intersection graph model $G_{n,m,p}$ for a certain range of m and p. Experimental results indicate that even outside this range the algorithm performs well, for example when $\alpha > 1$. It is clear

that the reconstruction of feature cliques becomes impossible once they are no longer maximal, which seems to happen when p is of order $m^{-1/2}$. It would be interesting to prove that this (or a different) algorithm succeeds up to this point.

Finally we would like to thank the authors of [14,15,9] for generous access to their databases.

References

[1] E. R. Scheinerman, Random interval graphs, Combinatorica 8 (4) (1988) 357–371.

[2] J. A. Fill, E. R. Scheinerman, K. B. Singer-Cohen, Random intersection graphs when $m = \omega(n)$: An equivalence theorem relating the evolution of the $G(n, m, p)$ and $G(n, p)$ models, Random Structures and Algorithms 16 (2) (2000) 156–176.

[3] M. Karoński, E. R. Scheinerman, K. B. Singer-Cohen, On random intersection graphs: The subgraph problem, Combinatorics, Probability and Computing 8 (1999) 131–159.

[4] K. B. Singer, Random intersection graphs, Ph.D. thesis, John Hopkins University, Baltimore, Maryland (1995).

[5] D. Stark, The vertex degree distribution of random intersection graphs, Random Structures and Algorithms 24 (3) (2004) 249–258.

[6] M. Behrisch, Component evolution in random intersection graphs, preprint (November 2004).

[7] E. Godehardt, J. Jaworski, Two models of random intersection graphs and their applications, Electronic Notes in Discrete Mathematics 10.
URL
`http://www1.elsevier.com/gej-ng/31/29/24/49/27/61/endm10036.ps`

[8] M. E. J. Newman, S. H. Strogatz, D. J. Watts, Random graphs with arbitrary degree distributions and their applications, Physical Review E 64.
URL `http://aps.arxiv.org/abs/cond-mat/0007235/`

[9] J.-L. Guillaume, M. Latapy, Bipartite structure of all complex networks, Information Processing Letters 90 (2004) 215–221.

[10] M. R. Garey, D. S. Johnson, Computers and Intractability, W.H. Freeman and Company, 1979.

[11] S. Janson, T. Łuczak, A. Ruciński, Random Graphs, John Wiley & Sons, 2000.

[12] R. Govindan, H. Tangmunarunkit, SCAN+Lucent internet map from the ISI, http://www.isi.edu/div7/scan/mercator/maps.html (November 1999).

[13] R. Albert, H. Jeong, A.-L. Barabási, Database of self-organized networks, http://www.nd.edu/ networks/database/index.html.

[14] C. Frömmel, C. Gille, A. Goede, C. Gröpl, R. Preissner, S. Hougardy, T. Nierhoff, M. Thimm, Accelerating screening of 3D protein data with a graph theoretical approach, Bioinformatics 19 (18) (2003) 2442–2447.

[15] A. Goede, R. Preissner, S. Hougardy, M. Thimm, Comparison of 2D similarity and 3D superposition. application to searching a conformational drug database, Journal of Chemical Information and Computer Sciences 44 (2004) 1816–1822.