

Approximate Discovery of Random Graphs*

Thomas Erlebach[†] Alexander Hall[‡] Matúš Mihal'ák[†]

Abstract

In the layered-graph query model of network discovery, a query at a node v of an undirected graph G discovers all edges and non-edges whose endpoints have different distance from v . We study the number of queries at randomly selected nodes that are needed for approximate network discovery in Erdős-Rényi random graphs $G_{n,p}$. We show that a constant number of queries is sufficient if p is a constant, while $\Omega(n^\alpha)$ queries are needed if $p = n^\varepsilon/n$, for arbitrarily small $\varepsilon > 0$, where $\alpha > 0$ is a constant depending only on ε . Our proof of the latter result yields also a somewhat surprising result on pairwise distances in random graphs which may be of independent interest: We show that for a random graph $G_{n,p}$ with $p = n^\varepsilon/n$, for arbitrarily small $\varepsilon > 0$, in any constant cardinality subset of the nodes the pairwise distances are all identical with high probability.

1 Introduction

A fundamental problem in the study of complex networks is how to obtain accurate information about the topology of a network using a limited number of measurements or observations. For example, attempts to map the Internet can be based on traceroute experiments [13] or on the analysis of BGP routing tables [22]. A simplified theoretical model of such *network discovery* settings, the so-called *layered-graph query model*, has been introduced in [3]. The goal is to discover the edges and non-edges (for $u, v \in V$, we call $\{u, v\}$ a non-edge if it is not an edge of the graph) of an unknown graph or network $G = (V, E)$ using a minimum number of queries; a query at a node v reveals all edges and non-edges whose endpoints have different distance from v .

The layered-graph query model can be interpreted in the following way: A query at v yields the shortest-path subgraph rooted at v , i.e., the set of all edges on shortest paths between v and any other node. To see that this is equivalent to our definition (where a query yields all edges

*Work partially supported by European Commission - Fet Open project DELIS IST-001907 Dynamically Evolving Large Scale Information Systems, for which funding in Switzerland is provided by SBF grant 03.0378-1.

[†]Department of Computer Science, University of Leicester, University Road, Leicester LE1 7RH, UK. {te17, mm215}@mcs.le.ac.uk

[‡]Institute for Theoretical Computer Science, ETH Zurich, CH-8092 Zurich, Switzerland, alex.hall@inf.ethz.com

and non-edges between vertices of different distance from v), note that an edge connects two vertices of different distance from v if and only if it lies on a shortest path between v and one of these two vertices. Furthermore, the shortest-path subgraph rooted at v implicitly confirms the absence of all edges between vertices of different distance from v that are not part of the shortest-path subgraph. A real-life scenario where the shortest-path subgraph rooted at a node of the network can be determined arises as follows: With traceroute tools, one can determine the path that packets take in the Internet if they are sent from a node to some destination. If each traceroute experiment returns a random shortest path to the destination, this path would be part of the shortest-path subgraph. One could then use repeated traceroute experiments to all destinations to discover all edges of the shortest-path subgraph. Making a query at v would mean getting access to node v and running repeated traceroute experiments from v to all other nodes. If we assume that the cost of getting access to a node is much higher than that of running the traceroute-experiments, minimizing the number of queries is a meaningful goal.

In the off-line version of network discovery, the goal is to verify with as few queries as possible a given graph or network $G = (V, E)$. In this case we also speak of *network verification*.

Simulation experiments with (scale-free as well as Erdős-Rényi) random graphs reported in [4] indicate that the number of queries needed to discover all edges and non-edges typically grows with the size of the graph, as expected, but in some cases appears to be bounded by a small constant independent of the size of the graph if only a large fraction (say, 95%) of the edges and of the non-edges needs to be discovered. This shows that for the practically relevant goal of approximate network discovery, a surprisingly small number of queries is often sufficient. Motivated by this experimental result, we now study this phenomenon analytically for Erdős-Rényi random graphs $G_{n,p}$. These are graphs on n nodes in which each possible edge is present independently with probability p . We consider the simple query strategy that selects the query nodes uniformly at random. We say that a set of random queries approximately discovers $G_{n,p}$ in expectation, if the expected number of edges discovered by the queries is at least a ρ -fraction of all edges, and the analogous condition is satisfied for non-edges. Here, ρ is a constant such as 0.95.

Surprisingly, we find that if p is a constant strictly between 0 and 1 (i.e., if we consider dense $G_{n,p}$ graphs), then a constant number of query nodes is sufficient to approximately discover $G_{n,p}$ in expectation, but if $p = n^\varepsilon/n$, for an arbitrarily small constant $\varepsilon > 0$, then $\Omega(n^\alpha)$ queries are necessary, where $\alpha > 0$ is a constant depending on ε . Our results show that the number of random queries needed to approximately discover $G_{n,p}$ depends on the density of the graph, and in the query model considered it is actually easier to discover dense random graphs than relatively sparse ones.

Related Work. There are several ongoing large-scale efforts to collect data representing local views of the Internet. Here we will only mention two. The most prominent one is probably the RouteViews project [22] by the University of Oregon. It collects data from a large number of so-called border gateway protocol routers. Essentially, for each router—which can be seen as a node in the Internet graph on the level of autonomous systems—its list of paths (to all other nodes in the network) is retrieved. More recently and, due to good publicity, very

successfully, the DIMES project [13] has started collecting data with the help of a volunteer community. Users can download a client that collects paths in the Internet by executing successive traceroute commands. A central server can direct each client individually by specifying which routes to investigate. Data obtained by these or similar projects has been used in heuristics to obtain maps of the Internet, basically by simply overlaying possible paths found by the respective project. There is an extensive body of related work studying various aspects of this approach, see, e.g., [8, 13, 22, 17, 18, 15, 2, 24, 12, 1, 10, 11].

In [3, 4], the network discovery and verification problems are introduced and several results for the layered-graph query model are presented. It is shown that the network verification problem cannot be approximated within a factor of $o(\log n)$ unless $\mathcal{P} = \mathcal{NP}$, proving that an approximation algorithm from [20] (see below) is best possible, up to constant factors. A useful lower bound formula is given for the optimal number of queries of a graph. A discussion of simulation experiments for four different heuristic discovery strategies on various types of graphs, including several random graph models, can be found in [4]. Moreover, the on-line setting (network discovery) is studied and several lower and upper bounds on the competitive ratio are given. A number of results for both the on-line and off-line setting have also been derived for the much weaker *distance query model* [14, 4], in which a query at node v reveals only the distances to all other nodes.

It turns out that the network verification problem in the layered-graph query model has previously been considered as the problem of placing landmarks in graphs [20]. Here, the motivation is to place landmarks in as few vertices of the graph as possible in such a way that each vertex of the graph is uniquely identified by the vector of its distances to the landmarks. The smallest number of landmarks that are required for a given graph G is also called the *metric dimension* of G [19]. For a survey of known results, we refer to [7].

The problem of determining whether k landmarks suffice (i.e., of determining if the metric dimension is at most k) is long known to be \mathcal{NP} -complete [16]; the mentioned inapproximability of $o(\log n)$ [3] for the network verification problem transfers directly to the problem of minimizing the number of landmarks. In [20] it is shown that the problem admits an $O(\log n)$ -approximation algorithm based on SETCOVER. For trees, they show that the problem can be solved optimally in polynomial time. Furthermore, they prove that one landmark is sufficient if and only if G is a path, and discuss properties of graphs for which 2 landmarks suffice. They also show that if k landmarks suffice for a graph with n vertices and diameter D , we must have $n \leq D^k + k$. For d -dimensional hypercubes, it was shown in [23] (using an earlier result from [21] on a coin weighing problem) that the metric dimension is asymptotically equal to $2d/\log_2 d$. See also [6] for further results on the metric dimension of Cartesian products of graphs.

Our Contribution and Outline. In Section 2 we give some preliminary definitions concerning (random) graphs and the layered-graph query model of network discovery. The stated results in $G_{n,p}$ graphs are presented in Section 3. Our analysis for constant p in Section 3.1 is based on the observation that the probability that a query at node q discovers an edge or non-edge $\{u, v\}$ is at least $2p(1 - p)$, which is the probability that q is adjacent to one of u, v

but not the other.

For the case of $p = n^\varepsilon/n$, treated in Section 3.2, we use bounds from [9] on the size of the i -neighborhood and on the size of the i -th breadth-first search layer of a node in $G_{n,p}$, for arbitrarily small $\varepsilon > 0$ depending on i . These bounds allow us to show that for an edge or non-edge $\{u, v\}$, a query node q is very likely to have the same distance from u and v (and thus does not discover the edge or non-edge). We generalize this in Section 3.3 to obtain the following result: For a random graph $G_{n,p}$ with $p = n^\varepsilon/n$, in any constant cardinality subset of the nodes the pairwise distances are all identical, with high probability (w.h.p.).

2 Preliminaries

Graphs and Neighborhoods. With $G = (V, E)$ we denote an undirected graph with $|V| = n$ nodes. For two distinct nodes $u, v \in V$, we say that $\{u, v\}$ is an *edge* if $\{u, v\} \in E$ and a *non-edge* if $\{u, v\} \notin E$. The set of non-edges of G is denoted by \overline{E} . For $u, v \in V$, let $d(u, v)$ be the distance between the nodes u, v , i.e., the number of edges on a shortest path between u and v . For a graph G and a node $v \in V$, the set of nodes at distance i of v is denoted as the i -th layer: $\Gamma_i(v) = \{u \in V | d(v, u) = i\}$. We define the i -neighborhood $N_i(v) = \bigcup_{j=0}^i \Gamma_j(v)$ to be the set of nodes within distance i of v .

$G_{n,p}$ denotes an Erdős-Rényi random graph on n nodes in which a pair of nodes appears as an edge with probability p .

The Layered-Graph Query Model. A *query* is specified by a node $v \in V$ and is called a query *at* v or simply the query v . The answer of a query at v consists of a set E_v of edges and a set \overline{E}_v of non-edges. These sets are determined as follows. Let E_v be the set of all edges connecting vertices in different layers (from v), and \overline{E}_v be the set of all non-edges whose endpoints are in different layers. Because the query result can be seen as a layered graph, we refer to this query model as the *layered-graph query model*.

A set $Q \subseteq V$ of queries discovers (all edges and non-edges of) a graph $G = (V, E)$, if $\bigcup_{q \in Q} E_q = E$ and $\bigcup_{q \in Q} \overline{E}_q = \overline{E}$. In the off-line case, we also say “verifies” instead of “discovers”. The network verification problem is to compute, for a given network G , a smallest set of queries that verifies G . The network discovery problem is the on-line version of the network verification problem. Its goal is to compute a smallest set of queries that discovers G . Here, the edges and non-edges of G are initially unknown to the algorithm, the queries are made sequentially, and the next query must always be determined based only on the answers of previous queries.

Discovering a Large Fraction of a Graph. Let $\rho \in (0, 1]$ be a constant, typically a “large” value close to 1. We say a query set $Q \subseteq V$ discovers a ρ -fraction of the graph, if $|\bigcup_{q \in Q} E_q| \geq \rho \cdot |E|$ and $|\bigcup_{q \in Q} \overline{E}_q| \geq \rho \cdot |\overline{E}|$.

Note that we require separately that a fraction of all edges and that a fraction of all non-edges should be discovered. This is important, since another seemingly natural definition which requires only that a fraction of all node pairs should be discovered, might be misleading for the

interesting case of sparse graphs. Here a query set discovering almost all non-edges but only some of the edges would be a valid solution, since the number of edges is small compared to the total number of node pairs. However, since only few edges were discovered, the resulting graph is far away from the actual one. This is avoided by the separate treatment of edges and non-edges.

For a random graph or if Q is a random variable, we say Q discovers a ρ -fraction of the graph in expectation, if $\mathbb{E} \left[\left| \bigcup_{q \in Q} E_q \right| \right] \geq \rho \cdot \mathbb{E} [|E|]$ and $\mathbb{E} \left[\left| \bigcup_{q \in Q} \overline{E}_q \right| \right] \geq \rho \cdot \mathbb{E} [|\overline{E}|]$.

3 Discovering a Large Fraction of a Random Graph

In this section we study the discovery strategy RANDOM which simply picks a given number k of query nodes at random from V (using the uniform distribution). We show that in a random graph $G_{n,p}$ already a constant number of such queries suffices to discover a ρ -fraction of the graph in expectation, if p is a constant.

Since one of the main motivations for studying the network discovery setting is to discover the Internet graph, the case of sparse graphs is practically more relevant. Interestingly, if $p = n^\varepsilon/n$ for certain arbitrarily small choices of $\varepsilon > 0$, RANDOM needs at least $\Omega(n^\alpha \cdot \rho)$ queries to discover a ρ -fraction of the graph in expectation, where $\alpha > 0$ depends on ε .

3.1 The Case of Constant p

To prove that RANDOM discovers a ρ -fraction of the graph in expectation with only constantly many queries is straightforward. We start by showing a helpful lemma on queries and one node pair.

Lemma 1. *For a random graph $G_{n,p} = (V, E)$ and three distinct nodes $q, u, v \in V$, a query at q discovers the node pair u, v with probability at least $2 \cdot p \cdot (1 - p)$. The probability that k queries discover u, v is at least $x = 1 - (1 - 2 \cdot p \cdot (1 - p))^k$.*

Proof. We call a node $w \in V$ a *candidate*, if w is directly connected to v and not to u or directly connected to u and not to v . If the query node q is a candidate, it surely discovers the node pair u, v . This is independent of whether $\{u, v\} \in E$ or $\{u, v\} \in \overline{E}$. The probability of this event is $\Pr [q \text{ is candidate}] = 2 \cdot p \cdot (1 - p)$.

If we have several query nodes Q , the events “ q is candidate” for $q \in Q$ are independent, since for each q the event depends on two distinct edges. Thus the probability that at least one query in Q discovers u, v is at least $\Pr [Q \text{ contains candidate}] = 1 - \Pr [\text{no } q \in Q \text{ is candidate}] = 1 - (1 - 2 \cdot p \cdot (1 - p))^k$, where $k = |Q|$. \square

The desired result is a corollary of the following theorem.

Theorem 2. *To discover a ρ -fraction of a $G_{n,p}$ graph in expectation, the RANDOM strategy needs at most $k = \log(1 - \rho) / \log(1 - 2 \cdot p \cdot (1 - p))$ queries.*

Proof. By Lemma 1 we know that k queries Q discover a node pair $u, v \in V$ with probability at least $x = 1 - (1 - 2 \cdot p \cdot (1 - p))^k$. The expected number of edges discovered by Q can be computed as $\mathbb{E}[\text{edges discovered by } Q] = \sum_{\{u,v\} \in E} \Pr[u, v \text{ discovered by } Q] \geq x \cdot |E|$. Similarly we obtain $\mathbb{E}[\text{non-edges discovered by } Q] \geq x \cdot |\bar{E}|$. Setting $x = \rho$ and solving for k gives the stated result. \square

3.2 The Case of $p = n^\varepsilon/n$

Given an arbitrarily chosen constant $i \in \mathbb{N}$, in this entire section we set $\varepsilon = 3/(6 \cdot i + 5)$ and $p = n^\varepsilon/n$. By $\alpha, \beta, c > 0$ we always denote appropriately chosen constants, possibly depending on ε . Let $G_{n,p} = (V, E)$ be a random graph. By $U = \{u_1, \dots, u_k\} \subset V$ we always denote an arbitrary node subset of constant cardinality. Let $N_i(U) := \bigcup_{\ell=1}^k N_i(u_\ell)$ denote the i -neighborhood of U . The event A plays a central role in our discussion and is defined as follows: for each $u \in U$ the size of its i -neighborhood is bounded from above by $|N_i(u)| \leq \bar{c} \cdot (np)^i$ and the size of its i -th layer is bounded from below by $|\Gamma_i(u)| \geq \underline{c}(np)^i$, for some constants $\bar{c}, \underline{c} > 0$. Additionally, there is no edge between the neighborhoods $N_i(u)$ and $N_i(v)$, for all pairs $u, v \in U$.

Lemma 3 states that event A holds w.h.p. Then in Lemma 4 we condition on event A and show that a node $w \in V \setminus N_i(U)$ is connected to two distinct i -th layers $\Gamma_i(u)$ and $\Gamma_i(v)$, for $u, v \in U$, with probability $\Omega(n^{-\beta})$, for a constant $\beta < 1$. We remark that the constant ε is chosen carefully on a “borderline”: small enough such that Lemma 3 still holds and large enough for Lemma 4 to hold for some constant $\beta < 1$. These two lemmata can be applied to prove that a query node q is at the same distance $2 \cdot (i + 1)$ from a node $u \in V$ and a node $v \in V$ w.h.p. Finally, we use this fact to show that $\Omega(n^\alpha \cdot \rho)$ queries are needed to discover a ρ -fraction of a $G_{n,p}$ graph in expectation, for some constant $\alpha > 0$. The proofs are based on two very helpful lemmata in [9] which give the tight bounds stated in event A on the size of the i -neighborhoods and the i -th layer.

Lemma 3. *Let i, ε, p be as given above. Let $G_{n,p} = (V, E)$ be a random graph and $U \subset V$ a constant cardinality node subset. Event A on $G_{n,p}$ and U holds with probability $1 - O(n^{-\alpha})$, for an appropriate constant $\alpha > 0$.*

Proof. We start by bounding the size of the neighborhoods and i -th layers. For a node $v \in V$ Lemma 2 from [9] states that $|N_i(v)| \leq \bar{c} \cdot (np)^i$ holds with probability at least $1 - o(n^{-1})$, for some constant $\bar{c} > 0$. To see that this bound actually holds simultaneously for all $u \in U$ with probability $1 - |U| \cdot o(n^{-1}) = 1 - o(n^{-1})$, simply consider the counter-events and apply the subadditivity of probabilities. Note that the cardinality $|U|$ is constant.

For a node $v \in V$ Lemma 8 from [9] states that if $G_{n,p}$ is connected, we have $|\Gamma_i(v)| \geq \underline{c}(np)^i$ with probability at least $1 - o(n^{-1})$, for some constant $\underline{c} > 0$. This bound actually holds simultaneously for all $u \in U$ with probability $1 - |U| \cdot o(n^{-1}) = 1 - o(n^{-1})$; again apply the subadditivity of probabilities to see this. Since $G_{n,p}$ is connected with probability at least $1 - o(n^{-1})$ for this range of p , cf. [5], the connectedness assumption can be dropped. In other

words, the bounds on the size of the i -th layers of all $u \in U$ hold for any $G_{n,p}$ with probability $1 - o(n^{-1})$.

Combining both the bounds for the neighborhoods and the bounds for the i -th layers of the nodes in U , we have shown that the first part of event A holds with probability $1 - o(n^{-1})$.

We now come to the second part of event A . Let $x = \bar{c} \cdot (np)^i$ and consider c arbitrary node subsets of cardinality at most x . The probability that there is no edge from one of these subsets to another is at least

$$(1-p)^{\binom{c}{2} \cdot x^2} \geq (1-p)^{c^2 \cdot \bar{c}^2 (np)^{2i}} \geq \exp\left(-\frac{p}{1-p} \cdot c^2 \cdot \bar{c}^2 (np)^{2i}\right) \geq \exp(-c' \cdot n^{\varepsilon-1} \cdot n^{2i\varepsilon})$$

for some constant $c' \geq c^2 \cdot \bar{c}^2 / (1-p)$. With $\alpha = -(\varepsilon - 1 + 2i\varepsilon) = 2/(6 \cdot i + 5)$ we get

$$\exp(-c' \cdot n^{\varepsilon-1} \cdot n^{2i\varepsilon}) \geq 1 - c' \cdot n^{-\alpha}.$$

We conclude that with probability $1 - c' \cdot n^{-\alpha} = 1 - O(n^{-\alpha})$ there is no edge between any of the c subsets of cardinality x . To see that this also holds for the constant number $c = |U|$ of neighborhoods $N_i(u)$ as long as $|N_i(u)| \leq x$, for $u \in U$, we consider neighborhoods in iteratively defined subgraphs of the original $G_{n,p}$. Instead of $N_i(u_1)$ consider the neighborhood $N_i^{(1)}(u_1)$ in the random graph $G^{(1)} = G_{n,p} \setminus (U \setminus \{u_1\})$. For $1 < j \leq k$ we iteratively define the random graphs $G^{(j)} = G_{n,p} \setminus (\bigcup_{\ell \in \{1, \dots, j-1\}} N_i^{(\ell)}(u_\ell) \cup (U \setminus \{u_j\}))$ for which u_j 's neighborhood is denoted by $N_i^{(j)}(u_j)$. By construction the neighborhoods $N_i^{(j)}(u_j)$ do not overlap and clearly $|N_i^{(j)}(u_j)| \leq |N_i(u_j)|$ holds, for $j \in \{1, \dots, k\}$. Moreover, by constructing $N_i^{(j)}(u_j)$ in such a way, no information about the edges between the individual $N_i^{(j)}(u_j)$ is revealed. Each such edge is still present with probability p . Therefore if $|N_i(u_j)| \leq x$, the computation goes through as above, yielding: with probability $1 - O(n^{-\alpha})$ there is no edge between any of the neighborhoods $N_i^{(j)}(u_j)$, $j \in \{1, \dots, k\}$. In this case we obviously have $N_i^{(j)}(u_j) = N_i(u_j)$.

Hence, the probability that the first and the second part of event A hold at the same time is at least $1 - o(n^{-1}) - O(n^{-\alpha}) = 1 - O(n^{-\alpha})$. Once more, this can be seen by considering the counter-events and applying the subadditivity of probabilities. \square

Lemma 4. *Let i, ε, p be as given above. Let $G_{n,p} = (V, E)$ be a random graph and $U \subset V$ a constant cardinality node subset. Conditioned on event A , a node $w \in V \setminus N_i(U)$ is connected to both $\Gamma_i(u)$ and $\Gamma_i(u')$ with probability $\Omega(n^{-\beta})$, for distinct $u, u' \in U$ and an appropriate constant $0 < \beta < 1$. This holds independently for all $w \in V \setminus N_i(U)$.*

Proof. Conditioning on event A reveals no information about the presence of edges between a node $w \in V \setminus N_i(U)$ and a node $v \in \Gamma_i(u)$, for $u \in U$. Such edges $\{w, v\}$ remain to be present independently with probability p .¹ Therefore, the probability that $w \in V \setminus N_i(U)$ is connected to both $\Gamma_i(u)$ and $\Gamma_i(u')$, for distinct $u, u' \in U$, is at least

$$(1 - (1-p)^{c(np)^i})^2 \geq (1 - \exp(-p \cdot c(np)^i))^2 \geq (1 - \exp(-c \cdot n^{\varepsilon-1+i\varepsilon}))^2.$$

¹Note on the other hand that by conditioning on event A we know that no edge between $w \in V \setminus N_i(U)$ and a node $v \in N_i(u) \setminus \Gamma_i(u)$, for $u \in U$, can be present. This will be used in the proof of Lemma 5.

With $a = -(\varepsilon - 1 + i\varepsilon) = (3i + 2)/(6i + 5)$ this gives

$$(1 - \exp(-\underline{c} \cdot n^{-a}))^2 \geq \left(\frac{\underline{c} \cdot n^{-a}}{1 + \underline{c} \cdot n^{-a}} \right)^2 \geq \Omega(n^{-2a}) \geq \Omega(n^{-\beta}),$$

for some constant β , with $2a \leq \beta < 1$. Since the edges considered for different w do not overlap, this holds independently for all $w \in V \setminus N_i(U)$. \square

Lemma 5. *Let i, ε, p be as given above. Let $G_{n,p} = (V, E)$ be a random graph and $q, u, v \in V$ three distinct nodes. A query at q discovers the node pair u, v with probability $O(n^{-\alpha})$, for an appropriate constant $\alpha > 0$.*

Proof. For the graph $G_{n,p}$ and $U = \{u, v, q\}$ we assume that event A holds and under this assumption show that w.h.p. $d(q, u) = d(q, v) = 2(i + 1)$. In the following we concentrate on $d(q, u)$. Let $V_{q,u} \subseteq V \setminus N_i(\{u, v, q\})$ be some constant fraction of all nodes, i.e., $n_{q,u} = |V_{q,u}| \geq n/c'$ for some constant $c' > 1$. This is possible, since by event A we know $|N_i(\{u, v, q\})| = o(n)$.

To show that $d(q, u) = 2(i + 1)$ w.h.p., it suffices to show that at least one of the nodes in $V_{q,u}$ is connected to both $\Gamma_i(q)$ and $\Gamma_i(u)$ w.h.p. Note that by construction no node in $V_{q,u}$ can be connected with a node in $N_i(q) \setminus \Gamma_i(q)$ or $N_i(u) \setminus \Gamma_i(u)$. The probability that at least one node in $V_{q,u}$ is connected to both $\Gamma_i(q)$ and $\Gamma_i(u)$ by Lemma 4 and with appropriate constants $c, c'', \alpha' > 0, \beta < 1$ is at least

$$1 - (1 - c \cdot n^{-\beta})^{n_{q,u}} \geq 1 - \exp(-c \cdot n^{-\beta} \cdot n_{q,u}) \geq 1 - \exp(-c/c' \cdot n^{1-\beta}) \geq 1 - \exp(-c''n^{\alpha'}),$$

With at least this probability $d(q, u) = 2(i + 1)$ holds. Note that by definition of $V_{q,u}$, there is still a constant fraction of all nodes left for $V_{q,v}$ and therefore $d(q, v) = 2(i + 1)$ holds w.h.p. as well.

Combining this with the probability for event A given by Lemma 3, we obtain that q is at the same distance from u and v with probability $1 - O(n^{-\alpha})$, for some constant $\alpha > 0$. Or equivalently: a query at q discovers the node pair u, v with probability at most $O(n^{-\alpha})$. \square

We are now ready to state our main theorem.

Theorem 6. *Let $i \in \mathbb{N}$ be a given constant and set $\varepsilon = 3/(6 \cdot i + 5)$, $p = n^\varepsilon/n$. To discover a ρ -fraction of a $G_{n,p}$ graph in expectation, the RANDOM strategy needs at least $\Omega(n^\alpha \cdot \rho)$ queries, for some appropriately chosen constant $\alpha > 0$.*

Proof. Assume we need $k = o(n^\alpha \cdot \rho)$ to discover a ρ -fraction in expectation. Let Q be a set of k queries returned by RANDOM. Then with Lemma 5 and $\mathbb{E}[|\overline{E}|] = \Omega(n^2)$ we get

$$\begin{aligned} \mathbb{E}[\text{non-edges discovered by } Q] &\leq \sum_{q \in Q} \sum_{u, v \in V: u \neq v} \Pr[u, v \text{ discovered by } q] \\ &\leq k \cdot O(n^{-\alpha}) \cdot n^2 = o(\rho) \cdot \mathbb{E}[|\overline{E}|]. \end{aligned}$$

This gives a contradiction and concludes the proof. \square

3.3 Distances Within a Constant Cardinality Subset of the Nodes

We generalize the result in Lemma 5 to the case of distances between the nodes of an arbitrary constant cardinality subset of V : all distances are identical w.h.p. for certain choices of ε and $p = n^\varepsilon/n$. We believe this property is interesting in itself, since it not necessarily seems intuitive at first sight. It obviously does not hold for “even sparser” graphs (e.g., $p = c/n$) or dense graphs ($p = c$).

Theorem 7. *With an arbitrarily chosen constant $i \in \mathbb{N}$, let $\varepsilon = 3/(6 \cdot i + 5)$ and $p = n^\varepsilon/n$. Let $G_{n,p} = (V, E)$ be a random graph and $U \subset V$ a constant cardinality node subset. All pairwise distances between the nodes in U are simultaneously equal to $2(i+1)$ with probability $1 - O(n^{-\alpha})$, for an appropriate constant $\alpha > 0$.*

Proof. We proceed as in Lemma 5, but instead of just two, we define $\binom{|U|}{2}$ disjoint sets $V_{u,v} \subseteq V \setminus N_i(U)$ with $|V_{u,v}| \geq n/c'$, for $u, v \in U$ and some constant $c' > 1$. Note that such a constant exists, since $|U|$ is constant. The argumentation goes through for each $V_{u,v}$ independently as above, giving the statement of the theorem. \square

4 Conclusion

We have introduced the notions of approximate network discovery and of discovering a large fraction of a graph in expectation. Motivated by previous computational experiments in random graphs, we have studied approximate network discovery in the $G_{n,p}$ model analytically for two different ranges of p . Surprisingly, we have been able to show that for constant p a constant number of queries suffices to discover a large fraction of a $G_{n,p}$ in expectation, whereas for certain small choices of p the number of queries needed grows with n .

The analysis of the latter case also gave an interesting result for constant cardinality subsets of the nodes of a $G_{n,p}$, for small p : w.h.p. all nodes of the subset are at exactly the same distance from each other.

It would be interesting to extend the analysis to other ranges of p . Analytically analysing network discovery in scale-free random graphs would be of interest as well, in particular due to the practical relevance. Many real world networks (e.g., the Internet graph or peer-to-peer networks) are believed to have scale-free properties.

Acknowledgments. The authors would like to thank Martin Marcinišzyn and Kostas Panagiotou for helpful discussions.

References

- [1] D. Achlioptas, A. Clauset, D. Kempe, and C. Moore. On the bias of traceroute sampling; or, power-law degree distributions in regular graphs. In *Proc. 37th STOC*, 2005.

- [2] P. Barford, A. Bestavros, J. Byers, and M. Crovella. On the marginal utility of deploying measurement infrastructure. In *Proc. ACM SIGCOMM Internet Measurement Workshop*, November 2001.
- [3] Z. Beerliová, F. Eberhard, T. Erlebach, A. Hall, M. Hoffmann, M. Mihal'ák, and L. S. Ram. Network discovery and verification. In *Proceedings of the International Workshop on Graph-Theoretic Concepts in Computer Science (WG'05)*, LNCS 3787, pages 127–138. Springer-Verlag, 2005.
- [4] Z. Beerliová, F. Eberhard, T. Erlebach, A. Hall, M. Hoffmann, M. Mihal'ák, and L. S. Ram. Network discovery and verification. *IEEE Journal on Selected Areas in Communications*, to appear.
- [5] B. Bollobás. *Random graphs*. Academic Press, New York, 1985.
- [6] J. Cáceres, C. Hernando, M. Mora, I. M. Pelayo, M. L. Puertas, C. Seara, and D. R. Wood. On the metric dimension of cartesian products of graphs. Manuscript, 2005.
- [7] G. Chartrand and P. Zhang. The theory and applications of resolvability in graphs: A survey. *Congr. Numer.*, 160:47–68, 2003.
- [8] B. Cheswick and H. Burch. Internet mapping project. <http://www.cs.bell-labs.com/who/ches/map/>.
- [9] F. Chung and L. Lu. The diameter of random sparse graphs. *Advances in Applied Math*, 26:257–279, 2001.
- [10] L. Dall'Asta, I. Alvarez-Hamelin, A. Barrat, A. Vázquez, and A. Vespignani. Statistical theory of internet exploration. *Phys. Rev. E*, 71, 2005.
- [11] L. Dall'Asta, I. Alvarez-Hamelin, A. Barrat, A. Vázquez, and A. Vespignani. Exploring networks with traceroute-like probes: theory and simulations. *Theoret. Comput. Sci.*, 355(1):6–24, April 2006.
- [12] G. Di Battista, T. Erlebach, A. Hall, M. Patrignani, M. Pizzonia, and T. Schank. Computing the types of the relationships between autonomous systems. *IEEE/ACM Transactions on Networking*, 2007. To appear.
- [13] DIMES. Mapping the Internet with the help of a volunteer community. <http://www.netdimes.org/>.
- [14] T. Erlebach, A. Hall, M. Hoffmann, and M. Mihal'ák. Network discovery and verification with distance queries. In *Proceedings of the 6th International Conference on Algorithms and Complexity (CIAC'06)*, LNCS 3998, pages 69–80. Springer, 2006.
- [15] L. Gao. On inferring autonomous system relationships in the Internet. *IEEE/ACM Trans. Networking*, 9(6):733–745, Dec 2001.
- [16] M. R. Garey and D. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. Freeman, 1979.
- [17] R. Govindan and A. Reddy. An analysis of internet inter-domain topology and route stability. In *Proc. IEEE INFOCOM*, April 1997.
- [18] R. Govindan and H. Tangmunarunkit. Heuristics for Internet map discovery. In *Proc. IEEE INFOCOM*, pages 1371–1380, March 2000.
- [19] F. Harary and R. Melter. The metric dimension of a graph. *Ars Combin.*, pages 191–195, 1976.
- [20] S. Khuller, B. Raghavachari, and A. Rosenfeld. Landmarks in graphs. *Discrete Appl. Math.*, 70:217–229, 1996.
- [21] B. Lindström. On a combinatorial detection problem I. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 9:195–207, 1964.
- [22] Oregon RouteViews. University of Oregon RouteViews project. <http://www.routeviews.org>.
- [23] A. Sebő and E. Tannier. On metric generators of graphs. *Math. Oper. Res.*, 29(2):383–393, 2004.
- [24] L. Subramanian, S. Agarwal, J. Rexford, and R. Katz. Characterizing the internet hierarchy from multiple vantage points. In *Proc. IEEE INFOCOM*, 2002.