Graph-Constrained Group Testing

Mahdi Cheraghchi, *Student Member, IEEE*, Amin Karbasi, *Student Member, IEEE*,
Soheil Mohajer, *Student Member, IEEE*, and Venkatesh Saligrama, *Senior Member, IEEE*

Abstract

Non-adaptive group testing involves grouping arbitrary subsets of n items into different pools. Each pool is then tested and defective items are identified. A fundamental question involves minimizing the number of pools required to identify at most d defective items. Motivated by applications in network tomography, sensor networks and infection propagation we formulate group testing problems on graphs. Unlike conventional group testing problems each group here must conform to the constraints imposed by a graph. For instance, items can be associated with vertices and each pool is any set of nodes that must be path connected. In this paper we associate a test with a random walk. In this context conventional group testing corresponds to the special case of a complete graph on n vertices.

For interesting classes of graphs we arrive at a rather surprising result, namely, that the number of tests required to identify d defective items is substantially similar to that required in conventional group testing problems, where no such constraints on pooling is imposed. Specifically, if $T(n)$ corresponds to the mixing time of the graph G, we show that with $m = O(d^2 T^2(n) \log(n/d))$ non-adaptive tests, one can identify the defective items. Consequently, for the Erdős-Rényi random graph $G(n, p)$, as well as expander graphs with constant spectral gap, it follows that $m = O(d^2 \log^3 n)$ non-adaptive tests are sufficient to identify d defective items. We next consider a specific scenario that arises in network tomography and show that $m = O(d^3 \log^3 n)$ non-adaptive tests are sufficient to identify d defective items. We also consider noisy counterparts of the graph constrained group testing problem and develop parallel results for these cases.

Mahdi Cheraghchi, Amin Karbasi and Soheil Mohajer are with the School of Computer and Communication Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland. Venkatesh Saligrama is with the Department of Electrical and Computer Engineering at Boston University, Boston, MA-02215. M. Cheraghchi is supported by an ERC Advanced Grant of A. Shokrollahi entitled “Error-correcting codes in science and communcation”. V. Saligrama is supported by the U.S. Department of Homeland Security under Award Number 2008-ST-061-ED0001 and NSF CAREER Award Number ECS 0449194. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security or the US National Science Foundation. Author names appear in alphabetical order.

I. INTRODUCTION

In this paper we introduce the graph constrained group testing problem motivated by applications in network tomography, sensor networks and infection propagation. While group testing theory (see [Dor43], [DH00] and more recently [AS09]), and its numerous applications, such as industrial quality assurance [SG59], DNA library screening [PL94], software testing [BG02], and multi-access communications [Wol85], have been systematically explored, the graph constrained group testing problem is new to the best of our knowledge.

Group testing involves identifying at most d defective items out of a set of n items. In non-adaptive group testing, which is the subject of this paper, we are given an $m \times n$ binary matrix, M, usually referred to as a test or measurement matrix. Ones on the jth row of M indicate which subset of the n items belong to the jth pool. A test is conducted on each pool; a positive outcome indicating that a defective item is part of the pool; and a negative test indicating that no defective items are part of the pool. The conventional group testing problem is to design a matrix M with minimum number of rows m that guarantees error free identification of the defective items. While the best known (probabilistic) pooling design requires a test matrix with $m = O(d^2 \log(n/d))$ rows, and an almost-matching lower bound of $m = \Omega(d^2 (\log n)/(\log d))$ is known on the number of pools (cf. [DH00, Chapter 7]), the size of the optimal test still remains open.

Note that in the standard group testing problem the test matrix M can be designed arbitrarily. In this paper we consider a generalization of the group testing problem to the case where the matrix M must conform to constraints imposed by a graph $G = (V, E)$. In general, as we will describe shortly, such problems naturally arise in several applications such as network tomography [Duf06], [NT07], sensor networks [NT06], and infection propagation [CHKV09]. While the graph constrained group testing problem has been alluded to in these applications the problem of test design or the characterization of the minimum number of tests to the best of our knowledge has not been addressed before. In this light our paper is the first to formalize the graph constrained group testing problem. In our graph group testing problem the n items are either vertices or links (edges) of the graph; at most d of them are defective. The task is to identify the defective vertices or edges. The test matrix M is constrained as follows: for items associated with vertices each row must correspond to a subset of vertices that are connected by a path on the graph; for items associated with links each row must correspond to links that are path connected in the line graph of G. The task is to design a $m \times n$ binary test matrix with minimum number of rows m that guarantees error free identification of the defective items.

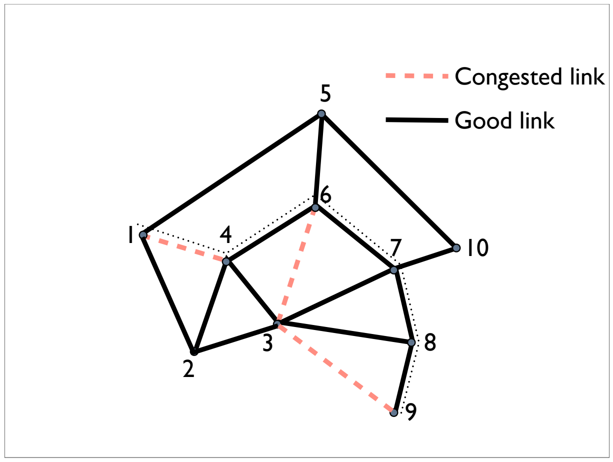

Fig. 1. The route $1 \to 4 \to 6 \to 7 \to 8 \to 9$ is valid while $2 \to 6 \to 5$ is not.

We will next describe several applications, which illustrate the graph constrained group testing problem.

A. *Network Tomography*

For a given network, identification of congested links from end-to-end path measurements is one of the key problems in network tomography [NT07], [Duf06]. In many settings of today's IP networks, there is one or a few links along the path which cause the packet losses in the path. Finding the locations of such congested links is sufficient for most of the practical applications.

This problem can be understood as a graph-constrained group testing as follows. We model the network as a graph $G = (V, E)$ where the set V denotes the network routers/hosts and the set E denotes the communication links (see Fig. 1). Suppose, we have a monitoring system that consists of one or more end hosts (so called *vantage* points) that can send and receive packets. Each vantage point sends packets through the network by assigning the routes and the end hosts.

All measurement results will be reported to a central server whose responsibility is to identify the congested links. Since the network is given, not any route is a valid route. A vantage point can only assign those routes which form a path in the graph G. The question of interest is to determine the number of measurements that is needed in order to identify the congested links in a given network.

B. *Sensor Networks*

The network tomography problem is further compounded in wireless sensor networks (WSN). As described in [NT06] the routing topology in WSN is constantly changing due to the inherent ad-hoc nature of the communication protocols. The sensor network is static with a given graph topology such as a geometric random graph. Sensor networks can be monitored passively or actively. In passive monitoring, at any instant, sensor nodes form a tree to route packets to the sink. The routing tree constantly changes unpredictably but must be consistent with the underlying network connectivity. A test is considered positive if the arrival time is significantly large, which indicates that there is a defective sensor node or a congested link. The goal is to identify defective links or sensor nodes based on packet arrival times at the sink. In active monitoring network nodes continuously calculate some high level, summarized information such as the average or maximum energy level among all nodes in the network. When the high level information indicates congested links, a low level and more energy consuming procedure is used to accurately locate the trouble spots.

C. Infection Propagation

Suppose that we have a large population where only a small number of people are infected by a certain viral sickness (e.g., a flu epidemic). The task is to identify the set of infected individuals by sending agents among them. Each agent contacts a pre-determined or randomly chosen set of people. Once an agent has made contact with an infected person, there is a chance that he gets infected, too. By the end of the testing procedure, all agents are gathered and tested for the disease. While this problem has been described in [CHKV09], the analysis ignores the inherent graph constraints that need to be further imposed. It is realistic to assume that, once an agent has contacted a person, the next contact will be with someone in close proximity of that person. Therefore, in this model we are given a random geometric graph that indicates which set of contacts can be made by an agent (see Fig. 2). Now, the question is to determine the number of agents that is needed in order to identify the set of infected people.

These applications present different cases where graph constrained group testing can arise. However there are important distinctions. In the wired network tomography scenario the links are the items and each row of the matrix M is associated with a route between any two vantage points. A test is positive if a path is congested, namely, if it contains at least one congested link. Note that in this case since the routing table is assumed to be static, the route between any two vantage points is fixed. Consequently, the matrix M is deterministic and the problem reduces to determining whether or not the matrix M satisfies identifiability.

Our problem is closer in spirit to the wireless sensor network scenario. In the passive case the links are the items and each row of the matrix M is associated with a route between a sensor node and the sink. A test is positive if a path is congested, namely, if it contains at least one congested link. Note that in

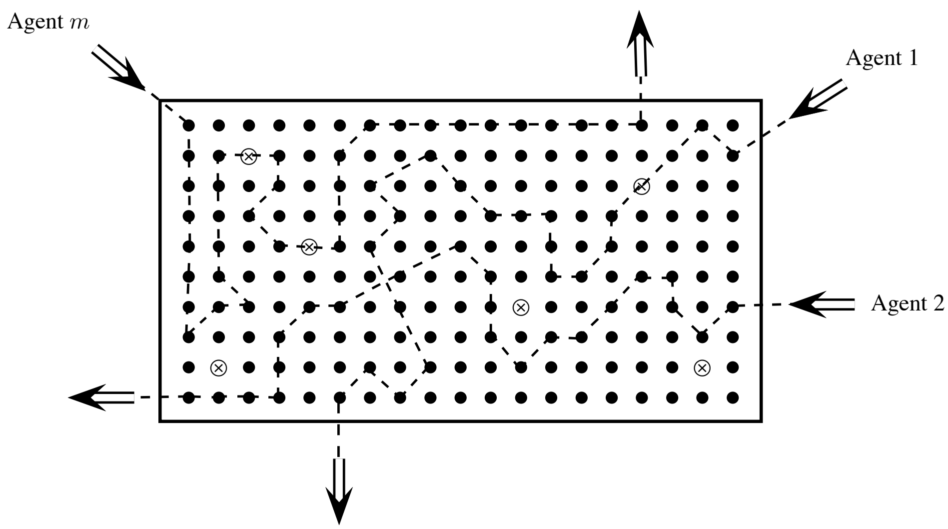

Fig. 2. Collective sampling using agents. The ⊗ symbols represent infected people whereas the healthy population is indicated by • symbols. The dashed lines show the group of people contacted by each agent [CHKV09].

this case since the routing table is constantly changing, the route between a sensor node and the sink is constantly changing as well. Nevertheless the set of possible routes must be drawn from the underlying connectivity graph. Consequently, the matrix M can be assumed to be random and the problem is to determine how many different tests are required to identify the congested links. Note that, in contrast to the wired scenario, tests conducted between the same sensor node and sink yields new information here. A similar situation arises in the active monitoring case as well. Here one could randomly query along different routes to determine whether or not a path is congested. These tests can be collated to identify congested links. Note that in the active case the test matrix M is amenable to design in that one could selectively choose certain paths over others by considering weighted graphs.

Motivated by the WSN scenario we describe pool designs based on random walks on graphs. As is well known a random walk is the state evolution on a finite reversible Markov chain. Each row of the binary test matrix is derived from the evolution of the random walk, namely, the ones on the jth row of M correspond to the vertices visited by the jth walk. This is close to the WSN scenario because as in the WSN scenario the path between two given nodes changes randomly. We develop several results in this context.

First, we consider random walks that start either at a random node or an arbitrary node but terminate in some appropriately chosen time t. By optimizing the length of the walk we arrive at a surprising result for interesting classes of graphs. Specifically we show that the number of tests required to identify d defective items is substantially similar to that required in conventional group testing problems, where no such constraints on pooling is imposed. The best known result for the number of tests required when no graphical constraints are imposed scales as $O(d^2 \log(n/d))$. For the graph constrained case we show that $m = O(d^2 T^2(n) \log(n/d))$ non-adaptive tests one can identify the defective items, where $T(n)$ corresponds to the mixing time of the graph G. Consequently, for the Erdős-Rényi random graph $G(n, p)$, as well as expander graphs with constant spectral gap, it follows that $m = O(d^2 \log^3 n)$ non-adaptive tests are sufficient to identify d defective items.

Next we consider unbounded-length random walks that originate at a source node and terminates at a sink node. Both the source node and the sink node can either be arbitrary or be chosen uniformly at random. This directly corresponds to the network tomography problem that arises in the WSN context. This is because the source nodes can be viewed as sensor nodes, while the sink node maybe viewed as the fusion center, where data is aggregated. At any instant, we can assume that a random tree originating at the sensor nodes and terminating at the sink is realized. While this random tree does not have cycles, there exist close connections between random walks and randomly generated trees. Indeed, it is well known that the so called loop-erased random walks, obtained by systematically erasing loops in random walks, to obtain spanning trees, is a method for sampling spanning trees from a uniform distribution [Wil96]. In this scenario we show that $m = O(d^3 \log^3 n)$ non-adaptive tests are sufficient to identify d defective items. By considering complete graphs we also establish that the cubic dependence on d in this result cannot be improved.

Finally, we also consider noisy counterparts of the graph constrained group testing problem and develop parallel results for these cases. Specifically, we consider the so called dilution model. In this case each item can be diluted in each test with some a priori known probability. This corresponds to the case when a test on a path with a congested link can turn out to be negative with some probability. We show that similar scaling result holds for this case as well.

Other group testing problems on graphs: Several variations of classical group testing have been studied in the literature that possess a graph theoretic nature. A notable example is the problem of learning hidden sparse subgraphs (or more generally, hypergraphs), defined as follows (cf. [Aig88]): Assume that, for a given graph, a small number of the edges are marked as defective. The problem is to use a small number of measurements of the following type to identify the set of defective edges: Each measurement specifies a subset of vertices, and the outcome would be positive iff the specified set contains a defective edge. Another variation concerns group testing with constraints defined by a rooted tree. Namely, the set of items correspond to the leaves of a given rooted tree, and each test is restricted to pool all the leaves that descend from a specified node in the tree (see [DH00, Chapter 12]). To the best of our knowledge, our work is the first variation to consider the natural restriction of the pools with respect to the *paths* on a given graph.

The rest of this paper is organized as follows. In Section II, we introduce our notation and mention some basic facts related to group testing and random walks on graphs. Section III formally describes the problem that we consider and states our main results. In Section IV we prove the main results, and finally, in Section V show instantiations of the result to the important cases of graph-constrained group testing on regular expander graphs and random graphs in the Erdős-Rényi model.

II. Definitions and Notation

In this section we introduce some tools, definition and notations which are used throughout the paper.

Definition 1. For two given boolean vectors S and T of the same length we denote their element-wise logical or by $S \vee T$. More generally, we will use $\bigvee_{i=1}^{d} S_i$ to denote the element-wise or of d boolean vectors. The logical subtraction of two boolean vectors $S = (s_1, \dots, s_n)$ and $T = (t_1, \dots, t_n)$, denoted by $S \setminus T$, is defined as a boolean vector which has a 1 at each position i if and only if $s_i = 1$ and $t_i = 0$. We also use $|S|$ to show the number of 1's in (i.e., the Hamming weight of) a vector S.

We often find it convenient to think of boolean vectors as characteristic vectors of sets. That is, $x \in \{0,1\}^n$ would correspond to a set $X \subseteq [n]$ (where $[n] := \{1, \dots, n\}$) such that $i \in X$ iff the entry at the ith position of x is 1. In this sense, the above definition extends the set theoretic notions of union, subtraction, and cardinality to boolean vectors.

Matrices that are suitable for the purpose of group testing are known as *disjunct* matrices. The formal definition is as follows:

Definition 2. An $m \times n$ boolean matrix M is called d-disjunct, if, for every column S_0 and every choice of d columns $S_1, \dots, S_d$ of M (different from S_0), there is at least one row at which the entry corresponding to S_0 is 1 and those corresponding to $S_1, \dots, S_d$ are all zeros. More generally, for an integer $e \geq 0$, the matrix is called (d, e)-disjunct if for every choice of the columns S_i as above, they satisfy

$$|S_0 \setminus \bigvee_{i=1}^{d} S_i| > e.$$

A $(d, 0)$-disjunct matrix is said to be d-disjunct.

A classical observation in group testing theory states that disjunct matrices can be used in non-adaptive group testing schemes to distinguish sparse boolean vectors (cf. [DH00]). More precisely, suppose that a d-disjunct matrix M with n columns is used as the measurement matrix; i.e., we assume that the rows of M are the characteristic vectors of the pools defined by the scheme. Then, the test outcomes obtained by applying the scheme on two distinct d-sparse vectors of length n must differ in at least one position. More generally, if M is taken to be (d, e)-disjunct, the test outcomes must differ in at least $e+1$ positions. Thus, the more general notion of (d, e)-disjunct matrices is useful for various "noisy" settings, where we are allowed to have a few false outcomes.

For our application, sparse vectors correspond to boolean vectors encoding the set of defective vertices, or edges, in a given undirected grph. Moreover, we aim to construct disjunct matrices that are also constrained to be *consistent* with the underlying graph.

Definition 3. Let $G = (V, E)$ be an undirected graph, and A and B be boolean matrices with $|V|$ and $|E|$ columns, respectively. The columns of A are indexed by the elements of V and the columns of B are indexed by elements of E. Then,

- The matrix A is said to be *vertex-consistent* with G if each row of A, seen as the characteristic vector of a subset of V, exactly represents the set of vertices visited by some walk on G.
- The matrix B is said to be *edge-consistent* with G if each row of B, seen as the characteristic vector of a subset of E, exactly corresponds to the set of edges traversed by a walk on G.

Note that the choice of the walk corresponding to each row of A or B need not be unique. Moreover, a walk may visit a vertex (or edge) more than once.

Definition 4. An undirected graph $G = (V, E)$ is called (D, c)-uniform, for some $c \geq 1$, if the degree of each vertex $v \in V$ (denoted by $\deg(v)$) is between D and cD.

Definition 5. The *point-wise distance* of two probability distributions μ, μ' on a finite space Ω is defined as

$$\|\mu - \mu'\|_\infty := \max_{i \in \Omega} |\mu(i) - \mu'(i)|,$$

where $\mu(i)$ (resp., $\mu'(i)$) denote the probability assigned by μ (resp., μ') to the outcome $i \in \Omega$. We say that the two distributions are δ-close if their point-wise distance is at most δ.

Definition 6. Let $G = (V, E)$ with $|V| = n$ be a (D, c)-uniform graph and denote by μ its stationary distribution. The δ*-mixing time* of G (with respect to the ℓ_∞ norm) is the smallest integer t such that a random walk of length t starting at any vertex in G ends up having a distribution μ' with $\|\mu' - \mu\|_\infty \leq \delta$. For concreteness, we define the quantity $T(n)$ as the δ-mixing time of G for $\delta := (1/2cn)^2$.

Throughout this work, the constraint graphs are considered to be (D, c)-uniform, for an appropriate choice of D and some (typically constant) parameter c. When $c = 1$, the graph is D-regular.

For a graph to have a small mixing time, a random walk starting from any vertex must quickly induce a uniform distribution on the vertex set of the graph. Intuitively this happens if the graph has no "bottle necks" at which the walk can be "trapped", or in other words, if the graph is "highly connected". The standard notion of *conductance*, as defined below, quantifies the connectivity of a graph.

Definition 7. Let $G = (V, E)$ be a graph on n vertices. For every $S \subseteq V$, define $\Delta(S) := \sum_{v \in S} \deg(v)$, $\bar{S} := V \setminus S$, and denote by $E(S, \bar{S})$ the number of edges crossing the cut defined by S and its complement. Then the *conductance* of G is defined by the quantity

$$\Phi(G) := \min_{S \subseteq V \colon \Delta(S) \leq |E|} \frac{E(S, \bar{S})}{\Delta(S)}.$$

Definition 8. Consider a particular random walk $W := (v_0, v_1, \ldots, v_t)$ of length t on a graph $G = (V, E)$, where the random variables $v_i \in V$ denote the vertices visited by the walk, and form a Markov chain. We distinguish the following quantities related to the walk W:

- For a vertex $v \in V$ (resp., edge $e \in E$), denote by π_v (resp., π_e) the probability that W passes v (resp., e).
- For a vertex $v \in V$ (resp., edge $e \in E$) and subset $A \subseteq V$, $v \notin A$ (resp., $B \subseteq E$, $e \notin B$), denote by $\pi_{v,A}$ (resp., $\pi_{e,B}$) the probability that W passes v but none of the vertices in A (resp., passes e but none of the edges in B).

Note that these quantities are determined by not only v, e, A, B (indicated as subscripts) but they also depend on the choice of the underlying graph, the distribution of the initial vertex v_0 and length of the walk t. However, we find it convenient to keep the latter parameters implicit when their choice is clear from the context.

In the previous definition, the length of the random walk was taken as a fixed parameter t. Another type of random walks that we consider in this work have their end points as a parameter and do not have an a priori fixed length. In the following we define similar probabilities related to the latter type of

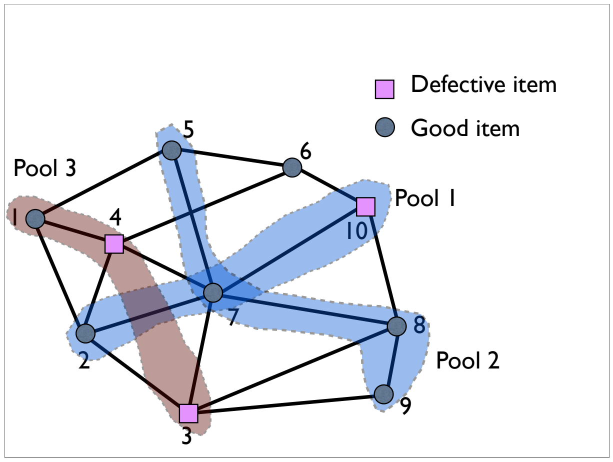

Fig. 3. The result of pool 1 is positive since it contains a defective item, whereas the result of pool 2 is negative since it does not contain a defective item. Pool 3 is not consistent with the graph and thus not allowed since the items are not connected by a path.

random walks.

Definition 9. Consider a particular random walk $W := (v_0, v_1, \ldots, u)$ on a graph $G = (V, E)$ that continues until it reaches a fixed vertex $u \in V$. We distinguish the following quantities related to W: For a vertex $v \in V$ (resp., edge $e \in E$) and subset $A \subseteq V$, $v \notin A$ (resp., $B \subseteq E$, $e \notin B$), denote by $\pi^{(u)}_{v,A}$ (resp., $\pi^{(u)}_{e,B}$) the probability that W passes v but none of the vertices in A (resp., passes e but none of the edges in B).

Again these quantities depend on the choice of G and the distribution of v_0 that we will keep implicit.

III. Problem Setting and Main Results

Problem Statement. Consider a given graph $G = (V, E)$ in which at most d vertices (resp., edges) are defective. The goal is to characterize the set of defective items using as small number of measurements as possible, where each measurements determines whether the set of vertices (resp., edges) observed along a path on the graph has a non-empty intersection with the defective set. We call the problem of finding defective vertices *vertex group testing* and that of finding defective edges *edge group testing*.

As mentioned earlier, not all sets of vertices can be grouped together, and only those that share a path on the underlying graph G can participate in a pool (see Fig. 3).

In the following, we introduce four random constructions (designs) for both problems. The proposed designs follow the natural idea of determining pools by taking random walks on the graph.

Design 1.

Given: a constraint graph $G = (V, E)$ with $r \geq 0$ designated vertices $s_1, \dots, s_r \in V$, and integer parameters m and t.

Output: an $m \times |V|$ boolean matrix M.

Construction: Construct each row of M independently as follows: Let $v \in V$ be any of the designated vertices s_i, or otherwise a vertex chosen uniformly at random from V. Perform a random walk of length t starting from v, and let the corresponding row of M be the characteristic vector of the set of vertices visited by the walk.

Design 2.

Given: a constraint graph $G = (V, E)$ and integer parameters m and t.

Output: an $m \times |E|$ boolean matrix M.

Construction: Construct each row of M independently as follows: Let $v \in V$ be any arbitrary vertex of G. Perform a random walk of length t starting from v, and let the corresponding row of M be the characteristic vector of the set of edges visited by the walk.

Design 3.

Given: a constraint graph $G = (V, E)$ with $r \geq 0$ designated vertices $s_1, \dots, s_r \in V$, a *sink node* $u \in V$, and integer parameter m.

Output: an $m \times |V|$ boolean matrix M.

Constructions: Construct each row of M independently as follows: Let $v \in V$ be any of the designated vertices s_i, or otherwise a vertex chosen uniformly at random from V. Perform a random walk starting from v until we reach u, and let the corresponding row of M be the characteristic vector of the set of vertices visited by the walk.

Design 4.

Given: a constraint graph $G = (V, E)$, a *sink node* $u \in V$, and integer parameter m.

Output: an $m \times |E|$ boolean matrix M.

Construction: Construct each row of M independently as follows: Let $v \in V$ be any arbitrary vertex of G. Perform a random walk, starting from v until we reach u, and let the corresponding row of M be the characteristic vector of the set of edges visited by the walk.

By construction, Designs 1 and 3 (resp., Designs 2 and 4) output boolean matrices that are vertex-

TABLE I

THE ASYMPTOTIC VALUES OF VARIOUS PARAMETERS IN THEOREM 11.

Parameter	Value
D_0	$O(c^2 d T^2(n))$
m_1, m_2	$O(c^4 d^2 T^2(n) \log(n/d))$
m_3	$O(c^8 d^3 T^4(n) \log(n/d))$
m_4	$O(c^9 d^3 D T^4(n) \log(n/d))$
t_1	$O(n/(c^3 d T(n)))$
t_2	$O(nD/(c^3 d T(n)))$
e_1, e_2, e_3, e_4	$\Omega(pd \log(n/d)/(1-p)^2)$
$m'_i, i \in [4]$	$O(m_i/(1-p)^2)$

(resp., edge-) consistent with the graph G. Our main goal is to show that, when the number of rows m is sufficiently large, the output matrices become d-disjunct (for a given parameter d) with overwhelming probability.

Remark 10. Designs 1 and 3 in particular provide two choices for constructing the measurement matrix M. Namely, the start vertices can be chosen within a fixed set of designated vertices, or, chosen randomly among all vertices of the graph. As we will see later, in theory, there is no significant difference between the two schemes. However, for some applications it might be the case that only a small subset of vertices are accessible as the starting points (e.g., in network tomography such a subset can be determined by the vantage points), and this can be modeled by an appropriate choice of the designated vertices in Designs 1 and 3.

The following theorem states the main result of this work.

Theorem 11. *Let $p \geq 0$ be a fixed parameter, and suppose that $G = (V, E)$ is a (D, c)-uniform graph on n vertices with δ-mixing time $T(n)$ (where $\delta := (1/2cn)^2$). Then there exist parameters with asymptotic values given in Table I such that, provided that $D \geq D_0$,*

1) *Design 1 with the path length $t := t_1$ and the number of measurements $m := m_1$ outputs a matrix M that is vertex-consistent with G. Moreover, once the columns of M corresponding to the designated vertices $s_1, \ldots, s_r$ are removed, the matrix becomes d-disjunct with probability $1-o(1)$. More generally, for $m := m'_1$ the matrix becomes (d, e_1)-disjunct with probability $1-o(1)$.*

2) *Design 2 with path length $t := t_2$ and $m := m_2$ measurements outputs a matrix M that is edge-consistent with G and is d-disjunct with probability $1 - o(1)$. More generally, for $m := m'_2$ the matrix becomes (d, e_2)-disjunct with probability $1 - o(1)$.*
3) *Design 3 with the number of measurements $m := m_3$ outputs a matrix M that is vertex-consistent with G. Moreover, once the columns of M corresponding to the designated vertices $s_1, \dots, s_r$ and the sink node u are removed, the matrix becomes d-disjunct with probability $1 - o(1)$. More generally, for $m := m'_3$ the matrix becomes (d, e_3)-disjunct with probability $1 - o(1)$.*
4) *Design 4 with the number of measurements $m := m_4$ outputs a matrix M that is edge-consistent with G and is d-disjunct with probability $1 - o(1)$. More generally, for $m := m'_4$ the matrix becomes (d, e_4)-disjunct with probability $1 - o(1)$.*

Remark 12. In Design 1, we need to assume that the designated vertices (if any) are not defective, and hence, their corresponding columns can be removed from the matrix M. By doing so, we will be able to ensure that the resulting matrix is disjunct. Obviously, such a restriction cannot be avoided since, for example, M might be forced to contain an all-ones column corresponding to one of the designated vertices and thus, fail to be disjunct.

Remark 13. By applying Theorem 11 on the complete graph (using Design 1), we get $O(d^2 \log(n/d))$ measurements, since in this case, the mixing time is $T(n) = 1$. Thereby, we recover the trade-off obtained by the probabilistic construction in classical group testing (note that classical group testing corresponds to graph-constrained group testing on the vertices of the complete graph).

We will show in Section V that, for our specific choice of $\delta := (1/2cn)^2$, the δ-mixing time of an Erdős-Rényi random graph $G(n, p)$ is (with overwhelming probability) $T(n) = O(\log n)$. This bound more generally holds for any graph with conductance $\Omega(1)$, and in particular, expander graphs with constant spectral gap. Thus we have the following result.

Theorem 14. *There is an integer $D_0 = \Omega(d \log^2 n)$ such that for every $D \geq D_0$ the following holds: Suppose that the graph G is either*

1) *A D-regular expander graph with normalized second largest eigenvalue (in absolute value) λ that is bounded away from 1; i.e., $\lambda = 1 - \Omega(1)$, or,*
2) *A random graph $G(n, D/n)$.*

Then for every $p \in [0, 1)$, with probability $1 - o(1)$ Designs 1, 2, 3, and 4 output (d, e)-disjunct matrices (not considering the columns corresponding to the designated vertices and the sink in Designs 1 and 3), for *some* $e = \Omega(dp \log n)$*, using respectively* m_1, m_2, m_3, m_4 *measurements, where* $m_1, m_2 = O(d^2 \log^3 n)$*,* $m_3 = O(d^3 \log^5 n)$*, and* $m_4 = O(d^3 D \log^5 n)$*.*

Example in Network Tomography. Here we illustrate a simple concrete example that demonstrates how our constructions can be used for network tomography in a simplified model. Suppose that a network (with known topology) is modeled by a graph with nodes representing routers and edges representing links that connect them, and it is suspected that at most d links in the network are congested (and thus, packets routed through them are dropped). Assume that, at a particular "source node" s, we wish to identify the set of congested links by distributing packets that originate from s in the network.

First, s generates a packet containing a time stamp t and sends it to a randomly chosen neighbor, who in turn, decrements the time stamp and forwards the packet to a randomly chosen neighbor, etc. The process continues until the time stamp reaches zero, at which point the packet is sent back to s along the same path it has traversed. This can be achieved by storing the route to be followed (which is randomly chosen at s) in the packet. Alternatively, for practical purposes, instead of storing the whole route in the packet, s can generate and store a random seed for a pseudorandom generator as a header in the packet. Then each intermediate router can use the specified seed to determine one of its neighbors to which the packet has to be forwarded.

Using the procedure sketched above, the source node generates a number of independent packets, which are distributed in the network. Each packet is either returned back to s in a timely manner, or, eventually do not reach s due to the presence of a congested link within the route. By choosing an appropriate timeout, s can determine the packets that are routed through the congested links.

The particular scheme sketched above implements our Design 2, and thus Theorem 11 implies that, by choosing the number of hops t appropriately, after generating a sufficient number of packets (that can be substantially smaller than the size of the network), s can determine the exact set of congested links. This result holds even if a number of the measurements produce false outcomes (e.g., a congested link may nevertheless manage to forward a packet, or a packet may be dropped for reasons other than congestion), in which case by estimating an appropriate value for the *noise parameter* p in Theorem 11 and increasing the number of measurements accordingly, the source can still correctly distinguish the congested links. Of course one can consider different schemes for routing the test packets. For example, it may be more desirable to forward the packets until they reach a pre-determined "sink node", an approach that is modeled by our Designs 3 and 4 above.

IV. PROOF OF THEOREM 11

Before discussing Theorem 11 and its proof, we introduce some basic propositions that are later used in the proof. The omitted proofs will be presented in the appendix. Throughout this section, we consider an underlying graph $G = (V, E)$ that is (D, c)-uniform, with mixing time $T(n)$ as in Definition 6.

Proposition 15. *Let $A, B_1, B_2, \ldots, B_n$ be events on a finite probability space, define $B := \cup_{i=1}^{n} B_i$, and suppose that:*

1) *For every $i \in [n]$, $\Pr[A \mid B_i] \leq \epsilon$.*
2) *for every set $S \subseteq [n]$ with $|S| > k$, $\cap_{i \in S} B_i = \emptyset$.*

Then, $\Pr[A \mid B] \leq \epsilon k$.

Proposition 16. *Denote by μ the stationary distribution of G. Then for each $v \in V$, $1/cn \leq \mu(v) \leq c/n$.*

Proposition 17. *For the quantities π_v and π_e in Definition 8, we have*

$$\pi_v = \Omega\left(\frac{t}{cnT(n)}\right), \quad \pi_e = \Omega\left(\frac{t}{cDnT(n)}\right).$$

In fact, a stronger statement than Proposition 17 can be obtained, that with noticeable probability, every fixed vertex (or edge) is hit by the walk at least once but not too many times nor too "early". This is made more precise in the following two propositions.

Proposition 18. *Consider any walk W in Design 1 (resp., Design 2). There is a $k = O(c^2 T(n))$ such that, for every $v \in V$ and every $e \in E$, the probability that W passes v (resp., e) more than k times is at most $\pi_v/4$ (resp., $\pi_e/4$).*

Proposition 19. *For any walk W in Design 1, let $v \in V$ be any vertex that is not among the designated vertices $s_1, \ldots, s_r$. Then the probability that W visits v within the first k steps is at most k/D.*

The following proposition shows that the distributions of two vertices on a random walk that are far apart from one another are almost independent:

Proposition 20. *Consider a random walk $w := (v_0, v_1, \ldots, v_t)$ on G starting from an arbitrary vertex, and suppose that $j \geq i + T(n)$. Let $\mathcal{E}$ denote any event that depends on the first i vertices on the walk. Then for every $u, v \in V$:*

$$|\Pr[v_i = u | v_j = v, \mathcal{E}] - \Pr[v_i = u | \mathcal{E}]| \leq 2/(3cn).$$

The following lemmas, which form the technical core of this work, lower bound the quantities $\pi_{v,A}$, $\pi_{e,B}$, $\pi^{(u)}_{v,A}$, $\pi^{(u)}_{e,B}$ as defined by Definitions 8 and 9.

Lemma 21. *There is a $D_0 = O(c^2 d T^2(n))$ and $t_1 = O(n/(c^3 d T(n)))$ such that whenever $D \geq D_0$, by setting the path lengths $t := t_1$ in Design 1 the following holds. Let $v \in V$, and $A \subseteq V$ be a set of at most d vertices in G such that $v \notin A$ and $A \cup \{v\}$ does not include any of the designated vertices $s_1, \ldots s_r$. Then*

$$\pi_{v,A} = \Omega\left(\frac{1}{c^4 d T^2(n)}\right). \tag{1}$$

Proof: Denote by μ the stationary distribution of G. We know from Proposition 16 that for each $u \in V$, $1/cn \leq \mu_u \leq c/n$.

Let $k = O(c^2 T(n))$ be the quantity given by Proposition 18, $\mathcal{B}$ denote the *bad event* that W hits some vertex in A. Moreover, let $\mathcal{G}$ denote the *good event* that W hits v no more than k times and never within the first $2T(n)$ steps. The probability of $\mathcal{G}$ is, by Propositions 18 and 19, at least

$$\Pr(\mathcal{G}) \leq 1 - 2T(n)/D - O(t/cnT(n)),$$

which can be made arbitrarily close to 1 (say larger than 0.99) by choosing D sufficiently large and t sufficiently small (as required by the statement). Now,

$$\begin{aligned} \pi_{v,A} &= \Pr[\neg\mathcal{B}, v \in W] \\ &\geq \Pr[\neg\mathcal{B}, v \in W, \mathcal{G}] \\ &= \Pr[v \in W, \mathcal{G}](1 - \Pr[\mathcal{B} \mid v \in W, \mathcal{G}]). \end{aligned} \tag{2}$$

By taking D large enough (in particular, $D = \Omega(c^2 d T^2(n))$), we can ensure that $2T(n)/D \leq \pi_v/4$. Combined with Proposition 18, we have $\Pr[v \in W, \mathcal{G}] \geq \pi_v/2$, and (2) gives

$$\pi_{v,A} \geq \pi_v(1 - \Pr[\mathcal{B} \mid v \in W, \mathcal{G}])/2. \tag{3}$$

Now we need to upperbound $\pi := \Pr[\mathcal{B} \mid v \in W, \mathcal{G}]$. Before doing so, fix some $i > 2T(n)$, and assume that $v_i = v$. Moreover, fix some vertex $u \notin A$ and assume that $v_0 = u$. We first try to upperbound $\Pr[\mathcal{B} \mid v_i = v, v_0 = u]$.

Let $\ell := i - T(n)$ and $\rho := i + T(n)$, and for the moment, assume that $T(n) + 1 < \ell < \rho < t$ (a "degenerate" situation occurs when this is not the case). Partition W into four parts:

$$\begin{aligned} W_1 &:= (v_0, v_1, \ldots, v_{T(n)}), \\ W_2 &:= (v_{T(n)+1}, v_{T(n)+2}, \ldots, v_{\ell-1}), \\ W_3 &:= (v_\ell, v_{\ell+1}, \ldots, v_r), \\ W_4 &:= (v_{r+1}, v_{\rho+2}, \ldots, v_t). \end{aligned}$$

For $j = 1, 2, 3, 4$, define

$$\pi_j := \Pr[W_j \text{ enters } A \mid v_i = v, v_0 = u].$$

Now we upperbound each of the π_j. In a degenerate situation, some of the W_i may be empty, and the corresponding π_j will be zero.

Each of the sub-walks W_2 and W_4 are "oblivious" of the conditioning on v_i and v_0 (because they are sufficiently far from both). In particular, the distribution of each vertex on W_4 is point-wise close to μ. Therefore, under our conditioning the probability that each such vertex belongs to A is at most $|A|(c/n+\delta) < 2dc/n$. The argument on W_2 is similar, but more care is needed. Without the conditioning on v_i, each vertex on W_2 has an almost-stationary distribution. Moreover, by Proposition 20, the conditioning on v_2 changes this distribution by up to $\delta' := 2/(3cn) < 1/n$ at each point. Altogether, for each $j \in \{T(n)+1, \ldots, \ell-1\}$, we have

$$\begin{aligned} \Pr[v_j \in A \mid v_i = v, v_0 = u] &\leq |A|(c/n + \delta + \delta') \\ &\leq 2dc/n. \end{aligned}$$

We conclude that $\pi_2 + \pi_4 \leq 2dct/n$.

In order to bound π_3, we observe that of all D or more neighbors of v_i, at most d can lie on A. Therefore,

$$\Pr[v_{i+1} \in A \mid v_i = v, v_0 = u] \leq d/D.$$

Similarly,

$$\Pr[v_{i+2} \in A \mid v_i = v, v_0 = u, v_{i+1}] \leq d/D,$$

regardless of v_{i+1} which means

$$\Pr[v_{i+2} \in A \mid v_i = v, v_0 = u] \leq d/D,$$

and so on. Similarly,

$$\Pr[v_{i-1} \in A \mid v_i = v] \leq d/D,$$

and by Proposition 20 (and time-reversibility), conditioning on v_0 changes this probability by at most $d\delta'$. Therefore,

$$\Pr[v_{i-1} \in A \mid v_i = v, v_0 = u] \leq d/D + d\delta',$$

and so on. Altogether, we get that

$$\pi_3 \leq 2dT(n)/D + 2T(n)/n.$$

Using the same reasoning, π_1 can be bounded as

$$\pi_1 \leq dT(n)/D + T(n)/n.$$

Finally, we obtain

$$\begin{aligned} \Pr[\mathcal{B} \mid v_i = v, v_0 = u] &\leq \pi_1 + \pi_2 + \pi_3 + \pi_4 \\ &\leq \frac{6dT(n)}{D} + \frac{2dct}{n}. \end{aligned} \tag{4}$$

The probability that the initial vertex is in A is at most d/n (as this happens only when the initial vertex is taken randomly), and by Proposition 20, conditioning on v_i changes this probability by at most $d\delta' < d/n$. Now we write

$$\begin{aligned} \Pr[\mathcal{B} \mid v_i = v] &\leq \Pr[v_0 \in A] + \Pr[\mathcal{B} \mid v_i = v, v_0 \notin A] \\ &\leq \Pr[v_0 \in A] + \pi_1 + \pi_2 + \pi_3 + \pi_4 \\ &\leq \frac{6dT(n)}{D} + \frac{4dct}{n}. \end{aligned}$$

Now, since $\Pr[\mathcal{G}]$ is very close to 1, conditioning on this event does not increase probabilities by much (say no more than a factor 1.1). Therefore,

$$\Pr[\mathcal{B} \mid v_i = v, \mathcal{G}] \leq 1.1 \left(\frac{6dT(n)}{D} + \frac{4dct}{n} \right).$$

Now in the probability space conditioned on $\mathcal{G}$, define events $\mathcal{G}_i$, $i = 2T(n) + 1, \ldots t$, where $\mathcal{G}_i$ is the event that $v_i = v$. Note that the intersection of more than k of the $\mathcal{G}_i$ is empty (as conditioning on $\mathcal{G}$ implies that the walk never passes v more than k times), and moreover, the union of these is the event that the walk passes v. Now we apply Proposition 15 to conclude that

$$\begin{aligned} \Pr[\mathcal{B} \mid v \in W, \mathcal{G}] &\leq 1.1k \left(\frac{6dT(n)}{D} + \frac{4dct}{n} \right) \\ &= O\left(c^2 T(n) \left(\frac{6dT(n)}{D} + \frac{4dct}{n} \right) \right) \end{aligned}$$

By taking $D = \Omega(c^2 dT^2(n))$ and $t = O(n/c^3 dT(n))$ we can make the right hand side arbitrarily small (say at most $1/2$). Now we get back to (3) to conclude, using Proposition 17, that

$$\pi_{v,A} \geq \pi_v/4 = \Omega\left(\frac{t}{cnT(n)}\right) = \Omega\left(\frac{1}{c^4 dT^2(n)}\right).$$

■

Similarly, we can bound the edge-related probability $\pi_{e,B}$ as in the following lemma. The proof of the lemma is very similar to that of Lemma 21, and is therefore skipped for brevity.

Lemma 22. *There is a $D_0 = O(c^2 dT^2(n))$ and $t_2 = O(nD/c^3 dT(n))$ such that whenever $D \geq D_0$, by setting the path lengths $t := t_2$ in Design 2 the following holds. Let $B \subseteq E$ be a set of at most d edges in G, and $e \in E$, $e \notin B$. Then*

$$\pi_{e,B} = \Omega\left(\frac{1}{c^4 dT^2(n)}\right). \tag{5}$$

In Designs 3 and 4, the quantities $\pi^{(u)}_{v,A}$ and $\pi^{(u)}_{e,B}$ defined in Definition 9 play a similar role as $\pi_{v,A}$ and $\pi_{e,B}$. In order to prove disjunctness of the matrices obtained in Designs 3 and 4, we will need lower bounds on $\pi^{(u)}_{v,A}$ and $\pi^{(u)}_{e,B}$ as well. In the following we show the desired lower bounds.

Lemma 23. *There is a $D_0 = O(c^2 dT^2(n))$ such that whenever $D \geq D_0$, in Design 3 the following holds. Let $v \in V$, and $A \subseteq V$ be a set of at most d vertices in G such that $v \notin A$ and $A \cup \{v\}$ is disjoint from $\{s_1, \dots s_r, u\}$. Then*

$$\pi^{(u)}_{v,A} = \Omega\left(\frac{1}{c^8 d^2 T^4(n)}\right). \tag{6}$$

Proof: Let D_0 and t_1 be quantities given by Lemma[1] 21. Let w_0 denote the start vertex of a walk performed in Design 3, and consider an infinite walk $W = (v_0, v_1, v_2, \dots)$ that starts from a vertex identically distributed with w_0. Let the random variables i, j, k respectively denote the time that W visits v, u, and any of the vertices in A for the first time. Therefore, $v_i = v$, $v_j = u$, and $v_k \in A$, $v_t \neq v$ for every $t < i$ and so on. Then the quantity $\pi^{(u)}_{v,A}$ that we wish to bound corresponds to the probability that $i < j < k$, that is, probability of the event that in W, the first visit of v occurs before the first visit of u and moreover, the first visit of u occurs before the first visit of any vertex in A. Observe that this event in particular contains the sub-event that $i \leq t_1$, $t_1 < j \leq 2t_1$, and $k > 2t_1$. Denote by $\mathsf{Good} \subseteq V^{t_1+1}$ the set of all sequences of $t_1 + 1$ vertices of G (i.e., walks of length t_1) that include v but not any of the vertices in $A \cup \{u\}$. Now, we can write

[1]In fact, as will be clear by the end of the proof, Lemma 21 should be applied with the sparsity parameter $d + 1$ instead of d. However, this will only affect constant factors that we ignore.

$$\begin{aligned}\pi^{(u)}_{v,A} &= \Pr[i<j<k] \geq \Pr[i \leq t_1 < j \leq 2t_1 < k] \\ &= \Pr[(i \leq t_1) \wedge (j > t_1) \wedge (k > t_1)] \Pr[t_1 < j \leq 2t_1 < k \mid (i \leq t_1) \wedge (j > t_1) \wedge (k > t_1)] \\ &= \Pr[(v_0, \ldots, v_{t_1}) \in \mathsf{Good}] \Pr[t_1 < j \leq 2t_1 < k \mid (v_0, \ldots, v_{t_1}) \in \mathsf{Good}] \qquad (7)\end{aligned}$$

The probability $\Pr[(v_0, \ldots, v_{t_1}) \in \mathsf{Good}]$ is exactly $\pi_{v,A\cup\{u\}}$ with respect to the start vertex w_0. Therefore, Lemma 21 gives the lower bound

$$\Pr[(v_0, \ldots, v_{t_1}) \in \mathsf{Good}] = \Omega\left(\frac{1}{c^4 d T^2(n)}\right).$$

Furthermore observe that, regardless of the outcome $(v_0, \ldots, v_{t_1}) \in \mathsf{Good}$, we have

$$\Pr[t_1 < j \leq 2t_1 < k \mid v_0, \ldots, v_{t_1}] = \pi_{u,A}$$

with respect to the start vertex v_{t_1}. Therefore, since $v_{t_1} \notin A \cup \{u\}$, again we can use Lemma 21 to conclude that

$$\Pr[t_1 < j \leq 2t_1 < k \mid (v_0, \ldots, v_{t_1}) \in \mathsf{Good}] = \Omega\left(\frac{1}{c^4 d T^2(n)}\right).$$

By plugging the bounds in (7) the claim follows. ■

A similar result can be obtained for Design 4 on the edges. Since the arguments are very similar, we only sketch a proof:

Lemma 24. *There is a $D_0 = O(c^2 d T^2(n))$ such that whenever $D \geq D_0$, in Design 4 the following holds. Let $B \subseteq E$ be a set of at most d edges in G, and $e \in E$, $e \notin B$. Then*

$$\pi^{(u)}_{e,B} = \Omega\left(\frac{1}{c^9 d^2 D T^4(n)}\right). \qquad (8)$$

Proof: (sketch) Similar to the proof of Lemma 23, we consider an infinite continuation $W = (v_0, v_1, \ldots)$ of a walk performed in Design 4 and focus on its first $t_1 + t_2$ steps, where t_1 and t_2 are respectively the time parameters given by Lemmas 21 and 22. Let

$$\begin{aligned} W_1 &:= (v_0, \ldots, v_{t_1}), \\ W_2 &:= (v_{t_1+1}, \ldots, v_{t_1+t_2}). \end{aligned}$$

Again following the argument of Lemma 23, we lower bound $\pi^{(u)}_{e,B}$ by the probability of a sub-event consisting the intersection of the following two events:

1) The event $\mathcal{E}_1$ that W_1 visits e but neither the sink node u nor any of the edges in B, and

2) The event $\mathcal{E}_2$ that W_2 visits the sink node u but none of the edges in B.

Consider the set $A \subseteq V$ consisting of the endpoints of the edges in B and denote by $v \in V$ any of the endpoints of e. Let $p := \pi_{v,A}$ (with respect to the start vertex v_0). Now, $\Pr[\mathcal{E}_1] \geq p/(cD)$ since upon visiting v, there is a $1/\deg(v)$ chance that the next edge taken by the walk turns out to be e. The quantity p in turn, can be lower bounded using Lemma 21. Moreover, regardless of the outcome of W_1, the probability that W_2 visits u but not B (and subsequently, the conditional probability $\Pr[\mathcal{E}_2 \mid \mathcal{E}_1]$) is at least the probability $\pi_{e',B}$ (with respect to the start vertex v_{t_1}), where $e' \in E$ can be taken as any edge incident to the sink node u. This latter quantity can be lower bounded using Lemma 22. Altogether, we obtain the desired lower bound on $\pi^{(u)}_{e,B}$. ■

Remark 25. It is natural to ask whether the exponent of d^2 in the denominator of the lower bound in Lemma 23 can be improved. We argue that this is not the case, by considering the basic where the underlying graph is the complete graph K_n and each walk is performed starting from a random node. Consider an infinite walk W starting at a random vertex, and the set of $d+2$ vertices $A' := A \cup \{u, v\}$. Due to the symmetry of the complete graph, we expect that the order at which W visits the vertices of A' for the first time is uniformly distributed among the $(d+2)!$ possible orderings of the elements of A'. However, in the event corresponding to $\pi^{(u)}_{v,A}$, we are interested in seeing v first, then u, and finally the elements of A. Therefore, for the case of complete graph we know that $\pi^{(u)}_{v,E} = O(1/d^2)$, and thus, the quadratic dependence on d is necessary even for very simple examples.

Remark 26. Another question concerns the dependence of the lower bound in Lemma 22 on the degree parameter D. Likewise Remark 25, an argument for the case of complete graph suggests that in general this dependence cannot be eliminated. For edge group testing on the complete graph, we expect to see a uniform distribution on the ordering at which we visit a particular set of edges in the graph. Now the set of edges of our interest consists of the union of the set $B \cup \{e\}$ and all the $n-1$ edges incident to the sink node u, and is thus of side $n+d$. The orderings that contribute to $\pi^{(u)}_{e,B}$ must have e as the first edge and an edge incident to u as the second edge. Therefore we get that, for the case of complete graph,

$$\pi^{(u)}_{e,B} = O(1/n) = O(1/D),$$

which exhibits a dependence on the degree in the denominator.

Now, we are ready to prove our main theorem.

Proof of Theorem 11: We prove the first part. For the other parts we need to follow the same reasoning. The high-level argument is similar to the well known probabilistic argument in classical group testing, but we will have to use the tools that we have developed so far for working out the details. By construction, the output matrix M is vertex-consistent with G. Now, take a vertex $v \in V$ and $A \subseteq V$ such that $v \notin A$, $|A| \leq d$, and $(\{v\} \cup A) \cap \{s_1, \dots, s_r\} = \emptyset$. For each $i = 1, \dots m_1$, define a random variable $X_i \in \{0, 1\}$ such that $X_i = 1$ iff the ith row of M has a 1 entry at the column corresponding to v and zeros at those corresponding to the elements of A. Let $X := \sum_{i=1}^{m_1} X_i$. Note that the columns corresponding to v and A violate the disjunctness property of M iff $X = 0$, and that the X_i are independent Bernoulli random variables. Moreover,

$$\mathbb{E}[X_i] = \Pr[X_i = 1] = \pi_{v,A},$$

since $X_i = 1$ happens exactly when the ith random walk passes vertex v but never hits any vertex in A. Now by an appropriate choice of D_0 and t_1 as stated in the lemma, we can ensure, using Lemma 21, that $\pi_{v,A} = \Omega(1/(c^4 d T^2(n)))$.

Denote by p_f the *failure probability*, namely that the resulting matrix M is not d-disjunct. By a union bound we get

$$\begin{aligned} p_f &\leq \sum_{v,A} (1 - \pi_{v,A})^{m_1} \\ &\leq \exp\left(d \log \frac{n}{d}\right) \cdot \left(1 - \Omega\left(\frac{1}{c^4 d T^2(n)}\right)\right)^{m_1}. \end{aligned}$$

Thus by choosing

$$m_1 = O\left(d^2 c^4 T^2(n) \log \frac{n}{d}\right)$$

we can ensure that $p_f = o(1)$, and hence, M is d-disjunct with overwhelming probability.

For the claim on (d, e_1)-disjunctness, note that a failure occurs if, for some choice of the columns, we have $X \leq e_1$. Set

$$\eta := p\pi_{v,A} = \Omega\left(\frac{p}{c^4 d T^2(n)}\right),$$

and $e_1 := \eta m_1'$. Now by a Chernoff bound, noting that $\mathbb{E}[X] = \pi_{v,A} m_1'$,

$$\begin{aligned} \Pr[X \leq \eta m_1'] &\leq \exp\left(-\frac{(\mathbb{E}[X] - \eta m_1)^2}{2\mathbb{E}[X]}\right) \\ &= \exp(-\mathbb{E}[X](1-p)^2/2). \end{aligned}$$

So now, by a union bound, the failure probability p_f becomes

$$p_f \leq \exp\left(d \log \frac{n}{d} - m_1'(1-p)^2 \pi_{v,A}/2\right),$$

which becomes $o(1)$ by choosing

$$m_1' = O\left(d^2 \log \frac{n}{d} c^4 T^2(n)/(1-p)^2\right).$$

■

V. Proof of Theorem 14

In Theorem 14 we consider two important instantiations of the result given by Theorem 11, namely when G is taken as an expander graph with constant spectral gap, and when it is taken as an Erdős-Rényi random graph $G(n,p)$. In the following we show that in both cases (and provided that p is not too small), the mixing time is $O(\log n)$. Then Theorem 11 will lead to the proof.

A. The Erdős-Rényi random graph

First, we present some tools for the case of random graphs. Consider a random graph $G(n,p)$ which is formed by removing each edge of the complete graph on n vertices independently with probability $1-p$. Our focus will be on the case where $np \gg \ln n$. In this case, the graph is (almost surely) connected and the degrees are highly concentrated around their expectations. In particular, we can show the following.

Proposition 27. *For every $\epsilon > 0$, with probability $1-o(1)$, the random graph $G(n,p)$ with $np \geq (2/\epsilon^2)\ln n$ is $(np(1-\epsilon), (1+\epsilon)/(1-\epsilon))$-uniform.*

Before we proceed, we need to bound the distance between the stationary distribution and the distribution obtained after t random steps on a graph. The following theorem, which is a direct corollary of a result in [SJ89], is the main tool that we will need.

Theorem 28. *Let G be an undirected graph with stationary distribution μ, and denote by $d_{\min}$ and $d_{\max}$ the minimum and maximum degrees of its vertices, respectively. Let μ' be the distribution obtained by any random walk on G in t steps. Then*

$$\|\mu' - \mu\|_\infty \leq (1-\Phi(G)^2/2)^t d_{\max}/d_{\min}.$$

In light of the above theorem, all we need to show is a lower bound on the conductance of a random graph. This is done in the following.

Lemma 29. *For every $\varphi < 1/2$, there is an $\alpha > 0$ such that a random graph $G = G(n,p)$ with $p \geq \alpha \ln n/n$ has conductance $\Phi(G) \geq \varphi$ with probability $1-o(1)$.*

Proof: First, note that by Proposition 27 we can choose α large enough so that with probability $1-o(1)$, the degree of each vertex in G is between $D(1-\epsilon)$ and $D(1+\epsilon)$, for an arbitrarily small $\epsilon > 0$ and $D := np$. We will suitably choose ϵ later.

Fix a set $S \subseteq V$ of size i. We wish to upper bound the probability that S makes the conductance of G undesirably low, i.e., the probability that $E(S, \bar{S}) < \varphi\Delta(S)$. Denote this probability by p_S. By the definition of conductance and (D, ϵ)-regularity of G, we only need to consider subsets of size at most ηn, for $\eta := (1+\epsilon)/2(1-\epsilon)$.

There are $i(n-i)$ "potential" edges between S and its complement in G, where each edge is taken independently at random with probability p. Therefore, the expected size of $E(S, \bar{S})$ is

$$\mu := Di(1 - i/n) \geq Di(1-\eta).$$

Now note that the event $E(S, \bar{S}) < \varphi\Delta(S)$ implies that

$$E(S, \bar{S}) < \varphi Di(1+\epsilon) < \varphi' \mu,$$

where $\varphi' := \varphi(1+\epsilon)/(1-\eta)$. So it suffices the upper bound the probability that $E(S, \bar{S}) < \varphi'\mu$. Note that, since $\varphi < 1/2$, we can choose ϵ small enough to ensure that $\varphi' < 1$. Now, by a Chernoff bound,

$$\begin{aligned} p_S &\leq \Pr[E(S, \bar{S}) < \varphi' \mu] \\ &\leq \exp(-(1-\varphi')^2 \mu) \\ &\leq n^{-i\alpha(1-\varphi')^2(1-\eta)}. \end{aligned}$$

Set α large enough (i.e., $\alpha \geq 2/(1-\varphi')^2(1-\eta)$) so that the right hand side becomes at most n^{-2i}. Therefore, with high probability, for our particular choice of S we have $E(S, \bar{S})/\varphi(S) \geq \varphi$.

Now we take a union bound on all possible choices of S to upper bound the probability of conductance becoming small as follows:

$$\begin{aligned} \Pr[\Phi(G) < \varphi] &\leq \sum_{i=1}^{\eta n} \binom{n}{i} n^{-2i} \\ &\leq \sum_{i=1}^{\eta n} n^{-i} = o(1). \end{aligned}$$

Thus with probability $1 - o(1)$, we have $\Phi(G) \geq \varphi$.

■

By combining Lemma 29 and Theorem 28 we get the following corollary:

Corollary 30. *There is an* $\alpha > 0$ *such that a random graph* $G = G(n,p)$ *with* $p \geq \alpha \ln n/n$ *has* δ*-mixing time bounded by* $O(\log(1/\delta))$ *with probability* $1 - o(1)$.

In particular, for our specific choice of $\delta := (1/2cn)^2$, the δ-mixing time of $G(n,p)$ would be $T(n) = O(\log n)$.

B. Expander Graphs with Constant Spectral Gap

A well known result in graph theory states that any regular graph with a normalized adjacency matrix whose second largest eigenvalue (in absolute value) is bounded away from 1 must have good expansion (cf. [HLW06]). As for regular graphs, the two notions of expansion and conductance (Definition 7) coincide, we can use Theorem 28 to obtain the following:

Lemma 31. *If* $G = (V, E)$ *is an expander graph with a (normalized) second largest eigenvalue that is bounded away from* 1 *by a constant, then* $T(n) = O(\log n)$.

We now have all the tools required for proving Theorem 14.

Proof of Theorem 14: Follows immediately by combining Theorem 11, Lemma 31, Corollary 30, and Proposition 27. ■

REFERENCES

[Aig88] M. Aigner. *Combinatorial Search*. Wiley-Teubner Series in Computer Science, Wiley, New York, 1988.

[AS09] G. Atia and V. Saligrama. Noisy group testing: An information theoretic perspective. In *Proceedings of the Allerton Conference on Communication, Control, and Computation, UIUC*, 2009.

[BG02] A. Blass and Y. Gurevich. Pairwise testing. *Bulletin of the EATCS*, 78:100–132, 2002.

[CHKV09] M. Cheraghchi, A. Hormati, A. Karbasi, and M. Vetterli. Compressed sensing with probabilistic measurements: A group testing solution. In *Proceedings of the Allerton Conference on Communication, Control, and Computation, UIUC*, 2009.

[DH00] D-Z. Du and F. Hwang. *Combinatorial Group Testing and its Applications*. World Scientific Publishing Company, second edition, 2000.

[Dor43] R. Dorfman. The detection of defective members of large populations. *Annals of Mathematical Statistics*, 14:436–440, 1943.

[Duf06] N. Duffield. Network tomography of binary network performance characteristics. *IEEE Transactions on Information Theory*, 52(12):5373–5388, 2006.

[HLW06] S. Hoory, N. Linial, and A. Wigderson. Expander graphs and their applications. *Bulletin of the AMS*, 43(4):439–561, 2006.

[NT06] H.X. Nguyen and P. Thiran. Using end-to-end data to infer lossy links in sensor networks. In *Proceedings of the* 25*th IEEE International Conference on Computer Communications (INFOCOM'06)*, 2006.

[NT07] H.X. Nguyen and P. Thiran. The boolean solution to the congested IP link location problem: Theory and practice. In *Proceedings of the* 26*th IEEE International Conference on Computer Communications (INFOCOM'07)*, pages 2117–2125, May 2007.

[PL94] P. Pevzner and R.J. Lipshutz. Towards DNA sequencing chips. In *Proceedings of MFCS*, volume 841 of *Lecture Notes in Computer Science*, pages 143–158, 1994.

[SG59] M. Sobel and P.A. Groll. Group testing to eliminate efficiently all defectives in a binomial sample. *Bell Systems Technical Journal*, 38:1179–1252, 1959.

[SJ89] A. Sinclair and M. Jerrum. Approximate counting, uniform generation and rapidly mixing Markov chains. *Information and Computation*, 82(1):93–133, 1989.

[Wil96] D. B. Wilson. Generating random spanning trees more quickly than the cover time. In *Proceedings of the Twenty-eighth Annual ACM Symposium on the Theory of Computing (Philadelphia, PA, 1996)*, pages 296–303, 1996.

[Wol85] J. Wolf. Born-again group testing: multiaccess communications. *IEEE Transactions on Information Theory*, 31:185–191, 1985.

VI. Appendix

A. Proof of Proposition 15

We can write

$$\begin{aligned} \Pr[A \mid B] &= \frac{\sum_{i=1}^{n} \Pr[A|B_i]\Pr[B_i]}{\Pr[B]} \\ &\leq \epsilon \cdot \frac{\sum_{i=1}^{n} \Pr[B_i]}{\Pr[B]} \\ &\leq \epsilon k. \end{aligned}$$

The last inequality is due to the fact that each element of the sample space can belong to at most k of the B_i, and thus, the summation $\sum_{i=1}^{n} \Pr[B_i]$ counts the probability of each element in B at most k times.

B. Proof of Proposition 16

A random walk on any graph graph G that is not bipartite converges to a stationary distribution μ, where

$$\mu(v) = \frac{d(v)}{2|E|}.$$

Since G is a (D, c)-uniform graph we know that $D \leq d(v) \leq cD$ and that $nD \leq 2|E| \leq ncD$.

C. *Proof of Proposition 17*

Let $t' := \lfloor t/T(n) \rfloor$, and for each $i \in \{0, \ldots, t'\}$, $w_i := v_{iT(n)}$. Denote by $W' := \{w_0, \ldots, w_{t'}\}$ a subset of of $t' + 1$ vertices visited by W. Obviously, π_v is at least the probability that $v \in W'$. Thus it suffices to lower bound the latter probability.

By the definition of mixing time, regardless of the choice of w_0, the distribution of w_1 is δ-close to the stationary distribution μ, which assigns a probability between $1/cn$ and c/n to v (by Proposition 16). Therefore, $\Pr[w_1 \neq v \mid w_0] \leq 1 - 1/cn + \delta$. Similarly, $\Pr[w_2 \neq v \mid w_0, w_1] \leq 1 - 1/2cn$, and so on. Altogether, this means that

$$\begin{aligned} \Pr[w_0 \neq v, w_1 \neq v, \ldots, w_{t'} \neq v] &\leq (1 - 1/cn + \delta)^{t/T(n)} \\ &\leq (1 - 1/2cn)^{t/T(n)} \\ &\leq \exp(-t/(2cnT(n))) \\ &\leq 1 - \Omega(t/(2cnT(n))). \end{aligned}$$

In the last equality we used the fact that $\exp(-x) \leq 1 - x/2$ for $0 \leq x \leq 1$. Thus the complement probability is lower bounded by $\Omega(t/(cnT(n))$. The calculation for π_e is similar.

D. *Proof of Proposition 18*

For every $i = 0, \ldots, t$, define a boolean random variable $X_i \in \{0, 1\}$ such that $X_i = 1$ iff $v_i = v$. Let $X := \sum_{i=0}^{t} X_i$ be the number of times that the walk visits v. For every $i \geq T(n)$, we have

$$\begin{aligned} \Pr(v_i = v) &= \mathbb{E}[X_i] \\ &\leq c/n + \delta \\ &\leq 2c/n. \end{aligned}$$

Define $X' := \sum_{i=T(n)}^{t} X_i$. By linearity of expectation, $\mathbb{E}[X'] < 2ct/n$, and by Markov's inequality,

$$\Pr[X' \geq \alpha c^2 T(n)] < \frac{2t}{\alpha cnT(n)}.$$

By taking α a large constant (e.g., $\alpha = 8$), and using Proposition 17, we can ensure that the bound on the probability is at most $\pi_v/4$. Thus the probability that $X \geq k$ for $k := (1 + \alpha c^2)T(n)$ is at most$\pi_v/4$. Proof for the edge case is similar.

E. Proof of Proposition 19

By the choice of v, the walk W visits v at the initial vertex v_0 if it starts at a random vertex. Thus the probability of visiting v at the initial step is $1/n \leq 1/D$.

Now, regardless of the outcome of the initial vertex v_0, the probability of visiting v as the second vertex v_1 is at most $1/D$, as v_0 has at least D neighbors and one is chosen uniformly at random. Thus, $\Pr[v_1 = v] \leq 1/D$, and similarly, for each i, $\Pr[v_i = v] \leq 1/D$. A union bound gives the claim.

F. Proof of Proposition 20

We can write

$$\Pr[v_i = u \mid v_j = v, \mathcal{E}] = \Pr[v_j = v \mid v_i = u, \mathcal{E}] \cdot \frac{\Pr[v_i = u \mid \mathcal{E}]}{\Pr[v_j = v \mid \mathcal{E}]}.$$

Now, from the definition of mixing time, we know that

$$|\Pr[v_j = v \mid v_i = u, \mathcal{E}] - \Pr[v_j = v \mid \mathcal{E}]| \leq 2\delta,$$

because regardless of the knowledge of $v_i = u$, the distribution of v_j must be δ-close to the stationary distribution. Therefore,

$$\begin{aligned} |\Pr[v_i = u | v_j = v, \mathcal{E}] - \Pr[v_i = u | \mathcal{E}]| &\leq 2\delta / \Pr[v_j = v \mid \mathcal{E}] \\ &\leq 2\delta/(1/cn - \delta) \leq 8\delta cn/3 \end{aligned}$$

G. Proof of Proposition 27

Let $\alpha := 6/\epsilon^2$ so that $np \geq \alpha \ln n$. Take any vertex v of the graph. The expected degree of v is np. As the edges are chosen independently, by a Chernoff bound, the deviation probability of $\deg(v)$ can be bounded as

$$\begin{aligned} \Pr[|\deg(v) - np| > \epsilon np] &\leq 2e^{-\epsilon^2 np/3} \\ &\leq 2n^{-\epsilon^2 \alpha/3}. \end{aligned}$$

This upper bounds the probability by $2n^{-2}$. Now we can use a union bound on the vertices of the graph to conclude that with probability at least $1 - 2/n$, the degree of each vertex in the graph is between $np(1 - \epsilon)$ and $np(1 + \epsilon)$.

H. Proof of Corollary 30

Choose α large enough so that, by Proposition 27 the graph becomes $(np(1-\epsilon),(1+\epsilon)/(1-\epsilon))$-uniform, for a sufficiently small ϵ and moreover, Lemma 29 can be applied to obtain $\Phi(G) = \Omega(1)$. Let μ' be the distribution obtained by any random walk on G in t steps and denote by μ the stationary distribution of G. Now Theorem 28 implies that,

$$\|\mu' - \mu\|_\infty \leq (1 - \Phi(G)^2/2)^t (1+\epsilon)/(1-\epsilon),$$

and thus, it suffices to choose $t = O(\log(1/\delta))$ to have $\|\mu' - \mu\|_\infty \leq \delta$.