


# Lower Bounds for Approximating Graph Parameters via Communication Complexity

Talya Eden<sup>1</sup>

School of Electrical Engineering, Tel Aviv University, Tel Aviv 6997801, Israel  
talyaa01@gmail.com

Will Rosenbaum

School of Electrical Engineering, Tel Aviv University, Tel Aviv 6997801, Israel  
will.rosenbaum@gmail.com

 <https://orcid.org/0000-0002-7723-9090>

---

## Abstract

In a celebrated work, Blais, Brody, and Matulef [7] developed a technique for proving property testing lower bounds via reductions from communication complexity. Their work focused on testing properties of functions, and yielded new lower bounds as well as simplified analyses of known lower bounds. Here, we take a further step in generalizing the methodology of [7] to analyze the query complexity of graph parameter estimation problems. In particular, our technique decouples the lower bound arguments from the representation of the graph, allowing it to work with any query type.

We illustrate our technique by providing new simpler proofs of previously known tight lower bounds for the query complexity of several graph problems: estimating the number of edges in a graph, sampling edges from an almost-uniform distribution, estimating the number of triangles (and more generally,  $r$ -cliques) in a graph, and estimating the moments of the degree distribution of a graph. We also prove new lower bounds for estimating the edge connectivity of a graph and estimating the number of instances of any fixed subgraph in a graph. We show that the lower bounds for estimating the number of triangles and edge connectivity also hold in a strictly stronger computational model that allows access to uniformly random edge samples.

**2012 ACM Subject Classification** Theory of computation → Lower bounds and information complexity

**Keywords and phrases** sublinear graph parameter estimation, lower bounds, communication complexity

**Digital Object Identifier** 10.4230/LIPIcs.APPROX-RANDOM.2018.11

**Related Version** A full version of this paper is available on arXiv [12], <https://arxiv.org/abs/1709.04262>.

**Acknowledgements** We thank Dana Ron and Oded Goldreich for their discussion and commentary on previous versions of this work. We also thank the anonymous reviewers for their valuable comments.

---

<sup>1</sup> This research was partially supported by a grant from the Blavatnik fund. The author is grateful to the Azrieli Foundation for the award of an Azrieli Fellowship, and for the Weinstein Institute for their support.



© Talya Eden and William B. Rosenbaum;

licensed under Creative Commons License CC-BY

Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2018).

Editors: Eric Blais, Klaus Jansen, José D. P. Rolim, and David Steurer; Article No. 11; pp. 11:1–11:18



Leibniz International Proceedings in Informatics

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

## 1 Introduction

Since the seminal work of Yao [34], (two party) communication complexity has become a central topic in computational complexity. While communication complexity is well-studied in its own right, its importance in complexity theory is highlighted by its numerous applications, for example, in proving lower bounds for Turing machines [29, 23], streaming algorithms [4, 6], circuit complexity [24], distributed algorithms [21, 31], learning theory [32], and algorithmic game theory [8, 19]. Lower bounds via reductions from communication complexity also tend to be conceptually simple: delicate analysis is delegated to a small handful of fundamental results in communication complexity.

In typical applications of communication complexity, instances of the problem considered can be readily partitioned into two or more pieces. This not the case in the context of property testing [16], where the goal is to distinguish instances of a problem that satisfy some property from those that are “far” from satisfying the property. Thus, the work of Blais, Brody, and Matulef [7] was surprising, as it drew a close connection between communication complexity and property testing. Specifically, in [7], the authors develop a framework for applying communication complexity to obtain lower bounds in the property testing model. Their methodology yielded new results in property testing, as well as simpler proofs of known results.

The framework of [7] (which was subsequently generalized by Goldreich in [15]) applied to the property testing of functions. In this context, an algorithm is given query access to a function, where each query consists of evaluating the function at a single input specified by the algorithm. The goal is to design algorithms that distinguish functions having some property – for example being monotonic, or  $k$ -linear – from those that are far from having the property in the sense that a constant fraction of the function’s outputs must be changed in order for it to satisfy the property. The basic methodology of [7] and [15] is as follows. Let  $P$  be a property,  $\Pi$  a two-party communication problem, and  $x$  and  $y$  the private inputs of the two parties. The idea is then to construct a function  $f_{x,y}$  with the following properties:

1. if  $\Pi(x, y) = 1$ , then  $f_{x,y}$  satisfies  $P$ ;
2. if  $\Pi(x, y) = 0$ , then  $f_{x,y}$  is far from satisfying  $P$ ;
3. for each  $z$  in the domain of  $f_{x,y}$ ,  $f_{x,y}(z)$  can be computed from  $x$  and  $y$  using at most  $B$  bits of communication.

The main result of [7] (cf. [15]) states that under the three conditions above, any algorithm that tests  $P$  requires at least  $R(\Pi)/B$  queries, where  $R(\Pi)$  is the randomized communication complexity of  $\Pi$ .

In this work, our goal is to prove lower bounds on the number of queries necessary to (approximately) solve various graph problems. In particular, we adapt the framework described above to the context of graph parameter estimation problems. Any graph  $G = (V, E)$  on  $n$  nodes can be viewed as a Boolean function whose values are the entries of the adjacency matrix of  $G$ . Thus we can directly apply the methodology of [7]. This view of graph property testing corresponds to the “dense graph” model introduced in [16]: the graph is accessed only through “pair queries” (i.e., asking if two nodes share an edge), and two graphs are far apart only if they differ on  $\epsilon n^2$  edges. However, this model is not suitable for analyzing graphs with  $o(n^2)$  edges. In order to deal with property testing and parameter estimation in non-dense families of graphs, more refined graph models were introduced. These models allow additional types of queries – degree queries and neighbor queries – that cannot be handled in any obvious way using the models of [7] and [15].

## 1.1 Our Results

Our main result (Theorem 3.3) gives a general reduction from communication complexity problems to graph problems. The theorem is closely related to Theorem 3.1 in [15], but it makes a further step at generalizing the results of [7]. In particular, our result makes no assumptions about the representation of the objects in question or the types of queries allowed (although nontrivial lower bounds are only obtained when the allowable queries can be efficiently simulated with a 2-party communication protocol). Thus, we believe our main result offers several advantages. Since our approach decouples the graph queries from the representation of the graph (e.g., by its adjacency lists or adjacency matrix), it can handle many different query models (possibly simultaneously). Further, our framework may be useful in distinguishing the relative power of different query access models. Finally, we believe our results unify and simplify previous lower bound arguments for graph parameter estimation (which typically relied upon careful analysis of statistical distances between families of graphs).

We apply our lower bound framework to the following graph problems

1. estimating the number of edges [14, 18] (Section 4.1),
2. sampling edges from an almost-uniform distribution [13] (Section 4.2),
3. estimating the number of triangles [9] (Section 4.3) and  $r$ -cliques [10] (see the full version [12])
4. estimating the moments of the degree distribution [11] (see the full version [12])
5. estimating the number of instances of any fixed subgraph  $H$  (Section 4.1),
6. estimating the edge-connectivity of a graph (Section 4.4).

The lower bounds for 1–4 match previously known results, while we believe 5 and 6 are new. In results 3 and 6, our lower bounds hold in a strictly stronger graph access model that additionally allows uniformly random edge samples. Interestingly, all of the lower bounds we prove are polynomial in the size of the instance, whereas the lower bounds presented in [7] and [15] are typically logarithmic in the instance size.

## 1.2 Related Work

A model for query-based sublinear graph algorithms was first presented in the seminal work of Goldreich, Goldwasser, and Ron [16] in the context of property testing. Their model is appropriate for dense graphs, as only “pair queries” (i.e., queries of the form “*Do vertices  $u$  and  $v$  share an edge?*”) are allowed. An analogous model for bounded degree graphs was introduced by Goldreich and Ron in [17]. In this model, it is assumed that all vertices have degree at most  $\Delta$ , and the graph is accessed via neighbor queries (“*Who is  $v$ 's  $i^{\text{th}}$  neighbor?*” for  $i \leq \Delta$ ). A similar model for sparse graphs was introduced by Parnas and Ron in [28], which does not assume that the maximum degree in the graph is bounded, and additionally allows degree queries (“*What is  $v$ 's degree?*”). Kaufman, Krivelevich and Ron [25] introduced the general graph model which allows all of the above queries – pair, degree and neighbor queries. The lower bounds we prove all apply to the general graph model.

The problem of estimating the average degree of a graph (or equivalently, the number of edges in a graph) was first studied by Feige [14]. In [14], Feige proves tight bounds on the number of degree queries necessary to estimate the average degree. In [18], Goldreich and Ron study the same problem, but in a model that additionally allows neighbor queries. In this model, they prove matching upper and lower bounds for the number of queries needed to estimate the average degree. In Corollary 4.2 we achieve the same lower bound as [18] for estimating the number of edges in a graph (their lower bound as well as ours also holds when

allowing for pair queries). The related problem of sampling edges from an almost-uniform distribution was recently studied by Eden and Rosenbaum in [13]. They prove tight bounds on the number of queries necessary to sample an edge in a graph from an almost-uniform distribution. In Theorem 4.6 we present a new derivation of the lower bound presented in [13].

Eden et al. [9] prove tight bounds on the number of queries needed to estimate the number of triangles. Their results were generalized by Eden et al. in [10] to approximating the number of  $r$ -cliques for any  $r \geq 3$ . In Corollary 4.3 and Theorem 4.7, we present a new derivation of the lower bound of [9] for estimating the number of triangles. In the full version [12], we generalize the triangle lower bound construction to obtain a lower bound for  $r$ -cliques matching that of [10]. The recent work of Bera and Chakrabarti [6] proves communication complexity lower bounds for counting subgraphs in a model where the graph’s edges are partitioned between two players. The authors apply this result to obtain *space* lower bounds in the streaming model, but the result also implies the query lower bound we achieve in Section 4.3 and uses similar techniques to our proof.<sup>2</sup>

In [20], Gonen et al. study the problem of approximating the number of  $s$ -star subgraphs, and give tight bounds on the number of (degree, neighbor, pair) queries needed to solve this problem. As noted by Eden et al. [11], counting  $s$ -stars is closely related to computing the  $s^{\text{th}}$  moment of the degree sequence. In [11], the authors provide a simpler optimal algorithm for computing the  $s^{\text{th}}$  moment of the degree sequence that has better or matching query complexity when the algorithm is also given an upper bound on the arboricity of the graph. In the full version of this paper [12], we prove the lower bounds of [11], which build upon and generalize the lower bounds of [20].

The recent paper of Aliakbarpour et al. [2] proposes an algorithm for estimating the number of  $s$ -star subgraphs that is allowed uniformly random edge samples as a basic query as well as degree queries. Interestingly, the additional computational power afforded by random edge queries allows their algorithm to break the lower bound of [20]. We remark that our lower bounds for estimating the number of triangles (Theorem 4.7) and edge connectivity (Theorem 4.9) still hold in this stronger query access model.

## 2 Preliminaries

### 2.1 Graph Query Models

Let  $G = (V, E)$  be a graph where  $n = |V|$  is the number of vertices and  $m = |E|$  is the number of edges. We assume that the vertices  $V$  are given distinct labels, say, from  $[n] = \{1, 2, \dots, n\}$ . For  $v \in V$ , let  $\Gamma(v)$  denote the set of neighbors of  $v$ , and  $\deg(v) = |\Gamma(v)|$  is  $v$ ’s degree. For each  $v \in V$ , we assume that  $\Gamma(v)$  is ordered by specifying some arbitrary bijection  $\Gamma(v) \rightarrow [\deg(v)]$  so that we may refer unambiguously to  $v$ ’s  $i^{\text{th}}$  neighbor. We let  $\mathcal{G}_n$  denote the set of all graphs on  $n$  vertices, together with all possible labelings of the vertices (from  $[n]$ ) and all orderings of the neighbors of each vertex, and we define  $\mathcal{G} = \bigcup_{n \in \mathbb{N}} \mathcal{G}_n$ .

We consider algorithms that access  $G$  via queries. In general, a **query** is an arbitrary function  $q : \mathcal{G} \rightarrow \{0, 1\}^*$ . We are interested in the following question: “Given a set  $Q$  of allowable queries and a graph problem  $g$  (e.g. a computing function, estimating a graph parameter, etc.), how many queries  $q \in Q$  are necessary to compute  $g$ ?”

---

<sup>2</sup> We thank an anonymous referee for bringing [6] to our attention.

Since we associate the vertex set  $V$  with the set  $[n]$ , we allow algorithms to have free access to the vertex set of the graph<sup>3</sup> – algorithms are only charged for obtaining information about the edges of a graph. We focus on models that allow the following types of queries:

1. **degree query**  $d : V \rightarrow [n - 1]$ , where  $d(v)$  returns  $v$ 's degree.
2. **neighbor query**  $\text{nbr}_i : V \rightarrow V \cup \{\emptyset\}$  for  $i \in [n - 1]$ , where  $\text{nbr}_i(v)$  returns  $v$ 's  $i^{\text{th}}$  neighbor if  $i \leq \deg(v)$  and  $\emptyset$  otherwise.
3. **pair query**  $\text{pair} : V \times V \rightarrow \{0, 1\}$ , where  $\text{pair}(u, v)$  returns 1 if  $(u, v) \in E$  and 0 otherwise  $(u, v) \notin E$ .

Taking  $Q$  to be the set of all neighbor, degree, and pair queries, we have  $|Q| = O(n^2)$ . This query model is known as the **general graph model** introduced in [25].<sup>4</sup> In Theorems 4.7 and 4.9, we also consider an expanded model that allows random edges to be sampled from a uniform distribution (cf. [2]).

We wish to characterize the **query complexity** of graph problems, that is, the minimum number of queries necessary to solve the problem. We consider randomized algorithms, and we assume the randomness is provided via a random string  $\rho \in \{0, 1\}^{\mathbf{N}}$ . Since the query complexity of the various estimation problems we consider depends on the measure being estimated, we use the expected query complexity, rather than the worst case query complexity.<sup>5</sup>

Most of our results are lower bounds on the number of expected queries necessary to estimate graph parameters.

► **Definition 2.1.** A **graph parameter** is a function  $g : \mathcal{G} \rightarrow \mathbf{R}$  that is invariant under any permutation of the vertices of each  $G \in \mathcal{G}$ . Formally,  $g$  is a graph parameter if for every  $n \in \mathbf{N}$ ,  $G = (V, E) \in \mathcal{G}_n$  and every permutation  $\pi : [n] \rightarrow [n]$ , the graph  $G_\pi = (V, E_\pi)$  defined by  $(v_{\pi(i)}, v_{\pi(j)}) \in E_\pi \iff (v_i, v_j) \in E$  satisfies  $g(G_\pi) = g(G)$ .

► **Definition 2.2.** Let  $g : \mathcal{G}_n \rightarrow \mathbf{R}$  be a graph parameter,  $\mathcal{A}$  an algorithm, and  $\varepsilon > 0$ . We say that  $\mathcal{A}$  **computes a (multiplicative)  $(1 \pm \varepsilon)$ -approximation** of  $g$  if for all  $G \in \mathcal{G}$ , the output of  $\mathcal{A}$  satisfies  $\Pr_\rho(|\mathcal{A}(G) - g(G)| \leq \varepsilon g(G)) \geq 2/3$ . Here, the probability is taken over the random choices of the algorithm  $\mathcal{A}$  (i.e., over the random string  $\rho$ ).

► **Remark 2.3.** In the general graph model, every graph  $G$  can be explored using  $O(\max\{n, m\})$  queries, for example, by using depth first search. Thus, we are interested in algorithms that perform  $o(\max\{n, m\})$  – or even better,  $o(n)$  – queries.

## 2.2 Communication Complexity Background

In this section, we briefly review some background on two party communication complexity and state a fundamental lower bound for the disjointness function. We refer the reader to [26] for a detailed introduction.

<sup>3</sup> In the sparse and general graph models of property testing, it is often assumed that the identities of the vertices are not known in advance. Rather, algorithms may sample vertices from a uniform distribution or discover new vertices that are neighbors of known vertices. Since the current paper aims to prove lower bounds, the assumption that the identities of vertices are known in advance is without loss of generality.

<sup>4</sup> Allowing only neighbor queries while assuming the graph has maximal degree  $d$ , and considering the distance with respect to  $n \cdot d$  is known as the “bounded degree graph” model, while only allowing pair queries and considering the distance with respect to  $n^2$  is the “dense graph” model.

<sup>5</sup> In communication complexity, it is customary to use worst-case complexity when analyzing randomized protocols. This is done without loss of generality, as protocol with a given expected communication cost can be converted to a protocol with asymptotically equal worst-case communication cost and slightly higher error probability. Such a transformation from expected to worst-case lower bounds is not generally possible for query complexity when the expected cost of a protocol depends on the parameter being estimated.

We consider two party communication complexity in the following setting. Let  $f : \{0, 1\}^N \times \{0, 1\}^N \rightarrow \{0, 1\}$  be a Boolean function. Suppose two parties, traditionally referred to as Alice and Bob, hold  $x$  and  $y$ , respectively, in  $\{0, 1\}^N$ . The (*randomized*) *communication complexity*<sup>6</sup> of  $f$  is the minimum number of bits that Alice and Bob must exchange in order for both of them to learn the value  $f(x, y)$ .

More formally, let  $\Pi$  be a communication protocol between Alice and Bob. We assume that  $\Pi$  is randomized, and that Alice and Bob have access to a shared random string,  $\rho$ . We say that  $\Pi$  **computes**  $f$  if for all  $x, y \in \{0, 1\}^N$ ,  $\Pr_\rho[\Pi(x, y) = f(x, y)] \geq 2/3$ , where the probability is taken over all random strings  $\rho$ . For fixed inputs  $x, y \in \{0, 1\}^N$  and random string  $\rho$ , we denote the number of bits exchanged by Alice and Bob using  $\Pi$  on input  $(x, y)$  and randomness  $\rho$  by  $|\Pi(x, y; \rho)|$ . The (**expected**) **communication cost**<sup>7</sup> of  $\Pi$  is defined by  $\text{cost}(\Pi) = \sup_{x, y} \mathbf{E}_\rho(|\Pi(x, y; \rho)|)$ . Finally, the (**randomized**) **communication complexity** of  $f$ , denoted  $R(f)$ , is the minimum cost of any protocol that computes  $f$ :  $R(f) = \min \{\text{cost}(\Pi) \mid \Pi \text{ computes } f\}$ .

The notion of communication complexity extends to partial functions in a natural way. That is, we may restrict attention to particular inputs for  $f$  and allow  $\Pi$  to have arbitrary output for all other values. Formally, we model this extension to partial functions via **promises** on the input of  $f$ . Let  $P \subseteq \{0, 1\}^N \times \{0, 1\}^N$ . We say that a protocol  $\Pi$  computes  $f$  for the promise  $P$  if for all  $(x, y) \in P$ ,  $\Pr_\rho(\Pi(x, y) = f(x, y)) \geq 2/3$ . The communication complexity of a promise problem (or equivalently, a partial function) is defined analogously to the paragraph above.

One of the fundamental results in communication complexity is a linear lower bound for the communication complexity of the disjointness function. Suppose Alice and Bob hold subsets  $A, B \subseteq [N]$ , respectively. The disjointness function takes on the value 1 if  $A \cap B = \emptyset$ , and 0 otherwise. By associating  $A$  and  $B$  with their characteristic vectors in  $\{0, 1\}^N$  (i.e.  $x_i = 1 \iff i \in A$  and  $y_j = 1 \iff j \in B$ ), we can define the disjointness function as follows.

► **Definition 2.4.** For any  $x, y \in \{0, 1\}^N$ , the **disjointness function** is defined by the formula  $\text{disj}(x, y) = \neg \bigvee_{i=1}^N x_i \wedge y_i$ .

The following lower bound for the communication complexity of  $\text{disj}$  was initially proved by Kalyansundaram and Schintger [22] with subsequent simplified proofs by Razborov [30] and Bar-Yossef et al. [5]. All of the results we present rely upon this fundamental lower bound.

► **Theorem 2.5** ([22, 30]). *The randomized communication complexity of the disjointness function is  $R(\text{disj}) = \Omega(N)$ . This result holds even if  $x$  and  $y$  are promised to satisfy  $\sum_{i=1}^N x_i y_i \in \{0, 1\}$  – that is, Alice’s and Bob’s inputs are either disjoint or intersect on a single point.*

The promise in Theorem 2.5 is known as **unique intersection**. We will also use a variant of the unique intersection problem that we refer to as the *k-intersection problem*.

<sup>6</sup> Throughout this paper, all algorithms and protocols are assumed to be randomized.

<sup>7</sup> It is more common in the literature to define the cost in terms of the worst case random string  $\rho$  rather than expected. However, for our purposes (since we consider the expected query complexity) it will be more convenient to use expected cost. We allow our protocols to err with (small) constant probability, so this difference only affects the communication complexity by a constant factor.



► **Definition 2.6.** Let  $x, y \in \{0, 1\}^N$ . We say that  $x$  and  $y$  are  **$k$ -intersecting** if  $\sum_{i=1}^N x_i y_i \geq k$ . The  $k$ -intersection function is defined by the formula  $\text{int}_k(x, y) = 1$  if  $\sum_i x_i y_i \geq k$  and 0 otherwise.

The following consequence of Theorem 2.5 is proven in Appendix A.

► **Corollary 2.7.**  $R(\text{int}_k) = \Omega(N/k)$ . The result holds even if  $x$  and  $y$  are promised to satisfy  $\sum_i x_i y_i \in \{0, k\}$ .

### 3 General Lower Bounds

In this section, we describe a framework for obtaining general query lower bounds from communication complexity. Let  $\mathcal{G}_n$  denote the family of graphs on the vertex set  $V = [n]$ , which we assume have labels 1 through  $n$ . We will use  $g : \mathcal{G}_n \rightarrow \{0, 1\}$  to denote a Boolean function on  $\mathcal{G}_n$ .

► **Definition 3.1.** Let  $P \subseteq \{0, 1\}^N \times \{0, 1\}^N$ . Suppose  $f : P \rightarrow \{0, 1\}$  is an arbitrary (partial) function, and let  $g$  be a Boolean function on  $\mathcal{G}_n$ . Let  $\mathcal{E} : \{0, 1\}^N \times \{0, 1\}^N \rightarrow \mathcal{G}_n$ . We call the pair  $(\mathcal{E}, g)$  an **embedding** of  $f$  if for all  $(x, y) \in P$  we have  $f(x, y) = g(\mathcal{E}(x, y))$ .

For a general embedding  $(\mathcal{E}, g)$  of a function  $f$ , the edges of  $\mathcal{E}(x, y)$  can depend on  $x$  and  $y$  in an arbitrary way. In order for the embedding to yield meaningful lower bounds, however, each allowable query  $q$  should be computable from  $x$  and  $y$  with little communication.

► **Definition 3.2.** Let  $q : \mathcal{G}_n \rightarrow \{0, 1\}^*$  be a query and  $(\mathcal{E}, g)$  an embedding of  $f$ . We say that  $q$  has **communication cost** at most  $B$  and write  $\text{cost}_{\mathcal{E}}(q) \leq B$  if there exists a (zero-error) communication protocol  $\Pi_q$  such that for all  $(x, y) \in P$  we have  $\Pi_q(x, y) = q(\mathcal{E}(x, y))$  and  $|\Pi_q(x, y)| \leq B$ .

► **Theorem 3.3.** Let  $Q$  be a set of allowable queries,  $f : P \rightarrow \{0, 1\}$ , and  $(\mathcal{E}, g)$  an embedding of  $f$ . Suppose that each query  $q \in Q$  has communication cost  $\text{cost}_{\mathcal{E}}(q) \leq B$ . Suppose  $\mathcal{A}$  is an algorithm that computes  $g$  using  $T$  queries (in expectation) from  $Q$ . Then the expected query complexity of  $\mathcal{A}$  is  $T = \Omega(R(f)/B)$ .

**Proof.** Suppose  $\mathcal{A}$  computes  $g$  using  $T$  queries in expectation. From  $\mathcal{A}$  we define a two party communication protocol  $\Pi_f$  for  $f$  as follows. Let  $x$  and  $y$  denote Alice and Bob's inputs, respectively, and  $\rho$  their shared public randomness. Alice and Bob both invoke  $\mathcal{A}$ , letting their shared randomness  $\rho$  be the randomness of  $\mathcal{A}$ . Whenever  $\mathcal{A}$  performs a query  $q$  that Alice or Bob cannot answer on their own, they communicate to the other party in order to determine the outcome of the query.<sup>8</sup> That is, they invoke  $\Pi_q$  in order to compute the response  $a$  to query  $q$ . The protocol terminates when  $\mathcal{A}$  halts and returns an answer  $\mathcal{A}(G)$ , at which point Alice and Bob determine their answer to  $f$  according to  $\mathcal{A}(G)$ .

Since  $\Pr_{\rho}(\mathcal{A}(G) = g(G)) \geq 2/3$ , and  $g(G) = f(x, y)$  it is clear that  $\Pi_f$  computes  $f$ . Further, since each  $\Pi_q$  satisfies  $|\Pi_q| \leq B(q)$ , we have  $\text{cost}(\Pi_f) = 2B \cdot T$ . Since  $\text{cost}(\Pi_f) \geq R(f)$ , we have  $T \geq R(f)/2B$ , as desired. ◀

Given the above Theorem, we suggest the following framework for proving graph query lower bounds.

<sup>8</sup> Since the randomness of  $\mathcal{A}$  is the shared randomness of Alice and Bob they both witness the same execution of  $\mathcal{A}$  and agree *without communication* on which query is being performed during each step. Further, since they both know the function  $\mathcal{E}$ , they can individually determine if a query  $q$  cannot be answered by the other party. In this case, they invoke  $\Pi_q$ .

1. Choose a “hard” communication problem  $f : P \rightarrow \{0, 1\}$ .
2. Define functions  $\mathcal{E} : P \rightarrow \mathcal{G}_n$  and  $g : \mathcal{G}_n \rightarrow \{0, 1\}$  such that  $(\mathcal{E}, g)$  is an embedding of  $f$  in the sense of Definition 3.1.
3. For each allowable query  $q \in Q$ , bound  $B$ , the number of bits that must be exchanged in order to simulate  $q$  given  $\mathcal{E}$ .

## 4 Lower Bounds for Particular Problems

In this section, we derive lower bounds for particular problems. In all cases, we allow  $Q$  to be the family of all degree, neighbor, and pair queries. In the case of the previously known lower bounds (estimating the number of edges, cliques,  $s^{\text{th}}$ -moment and sampling an edge from an almost uniform distribution) the graph constructions are similar or identical to the lower bound constructions in the original works. Our contribution is in the simplicity of the analysis.

### 4.1 Counting Subgraphs

Let  $G = (V, E)$  be a graph and  $H = (V_H, E_H)$  be a fixed graph with  $|V_H| = k$ . We denote the number of instances of  $H$  in  $G$  by  $h_H(G)$ . That is,  $h_H(G)$  is the number of subgraphs  $G' = (V', E')$  with  $V' \subseteq V$ ,  $E' \subseteq E$ , and  $|V'| = k$  such that  $G'$  is isomorphic to  $H$ .

► **Theorem 4.1.** *Let  $k$  be a fixed constant,  $H = (V_H, E_H)$  a fixed graph on  $k$  vertices ( $|V_H| = k$ ), and  $G'$  any graph on  $n/2$  vertices. For any  $\mu \leq \binom{n/2}{k}$ , any algorithm  $\mathcal{A}$  that distinguishes between graphs  $G$  on  $n$  vertices satisfying  $h_H(G) = h_H(G')$  and  $h_H(G) = h_H(G') + \mu$  requires  $\Omega(n/\mu^{1/k})$  degree, neighbor, or pair queries in expectation.*

**Proof.** We apply Theorem 3.3 with  $f = \text{disj}$ , the disjointness function with input size  $N = \Omega(n/\mu^{1/k})$ . For fixed  $n$  and  $x, y \in \{0, 1\}^N$  we construct a graph  $G = \mathcal{E}(x, y)$  on  $n$  vertices as follows. Take  $V = \{v_1, v_2, \dots, v_n\}$ . Let  $\ell$  be the smallest integer satisfying  $\binom{\ell}{k} \geq \mu$  so that  $\ell = O(\mu^{1/k})$ . Take  $N = n/2\ell$ . We partition the first  $n/2$  vertices into  $N$  sets of size  $\ell$  which we denote  $K_1, K_2, \dots, K_N$ . That is,  $K_j = \{v_{j(\ell-1)+1}, v_{j(\ell-1)+2}, \dots, v_{j\ell}\}$ . The set of edges within  $K_j$  is determined by  $x_j$  and  $y_j$ . If  $x_j = y_j = 1$ , then  $K_j$  is a clique. Otherwise,  $K_j$  is a set of isolated vertices. Formally,

$$\text{for all } u, v \in K_j, (u, v) \in E \iff x_j = y_j = 1. \quad (1)$$

Edges are added to the remaining  $n/2$  vertices of  $V$  (i.e., vertices  $V_2 = \{v_{n/2+1}, \dots, v_n\}$ ) so that the induced subgraph on  $V_2$  is isomorphic to  $G'$ . Finally, let  $g : \mathcal{G}_n \rightarrow \{0, 1\}$  be the (partial) function defined by

$$g(G) = \begin{cases} 1 & \text{if } h_H(G) \leq h_H(G') \\ 0 & \text{if } h_H(G) \geq h_H(G') + \mu \end{cases}$$

We claim that  $(\mathcal{E}, g)$  is an embedding of  $\text{disj}$ . To see this, first note that if  $\text{disj}(x, y) = 1$ , then the condition of Equation (1) is never satisfied. Thus,  $G$  is isomorphic to  $G'$ , plus  $n/2$  isolated vertices. In particular,  $h_H(G) = h_H(G')$ . On the other hand, if  $\text{disj}(x, y) = 0$ , then there exists  $j \in [N]$  with  $x_j = y_j = 1$ . Thus  $K_j$  is a clique on  $\ell$  vertices, implying that  $h_H(K_j) \geq \binom{\ell}{k} \geq \mu$ . Therefore,  $h_H(G) \geq h_H(G') + h_H(K_j) \geq h_H(G') + \mu$ , and the claim follows.

Finally, in order to apply Theorem 3.3, we must show that each degree, neighbor, or pair query can be simulated by Alice and Bob (who know  $x$  and  $y$ , respectively) using few bits. Let  $u, v \in V$ .



**Degree query.** Notice that if  $u \notin K_1 \cup \dots \cup K_N$ , then Alice and Bob can compute  $d(u)$  with no communication, as  $d(u)$  does not depend on  $x$  or  $y$ . If  $u \in K_j$ , then Alice and Bob can compute  $d(u)$  by exchanging  $x_j$  and  $y_j$ , which requires 2 bits.

**Neighbor query.** Again, if  $u \notin K_1 \cup \dots \cup K_n$ , Alice and Bob can compute  $\text{nbr}_i(u)$  without communication (by specifying some ordering on the edges of  $G'$  ahead of time). For  $u \in K_j = \{v_{\ell(j-1)+1}, \dots, v_{\ell j}\}$ , Alice and Bob can again compute  $\text{nbr}_i(u)$  by exchanging  $x_j$  and  $y_j$ . To this end, if  $x_j = 0$  or  $y_j = 0$ , then  $\text{nbr}_i(u) = \emptyset$  for all  $i$ . If  $x_j = y_j = 1$ , then we can order the neighbors of  $u = v_{\ell(j-1)+z}$  as follows: the  $i^{\text{th}}$  neighbor of  $u$  is  $v_{\ell(j-1)+z+i}$ , where the sum  $z+i$  is computed modulo  $\ell$ .

**Pair query.** The query  $\text{pair}(u, v)$  depends only on  $x$  and  $y$  if  $u, v \in K_j$  for some  $j$ . In this case,  $\text{pair}(u, v) = 1$  if and only if  $x_j = y_j = 1$ . Thus Alice and Bob can simulate  $\text{pair}(u, v)$  with 2 bits of communication.

Thus, all queries can be simulated using at most 2 bits of communication between Alice and Bob. Therefore, by Theorem 3.3, any algorithm that computes  $g$  requires  $\Omega(R(\text{disj})/B) = \Omega(N/2) = \Omega(n/\ell) = \Omega(n/\mu^{1/k})$  queries, as desired. ◀

► **Corollary 4.2** (Theorem 3.2 in [18]). *Suppose  $\mathcal{A}$  is an algorithm that gives a  $(1 + \varepsilon)$  multiplicative approximation to the number of edges in a graph using neighbor, degree, and pair queries. Specifically, for any  $\varepsilon > 0$ , on any input graph  $G = (V, E)$  with  $|V| = n$  and  $|E| = m$ ,  $\mathcal{A}$  outputs an estimate  $\hat{m}$  satisfying  $\Pr(|\hat{m} - m| < \varepsilon) \geq 2/3$ . Then the expected query complexity of  $\mathcal{A}$  is  $\Omega(n/\sqrt{\varepsilon m})$ .*

**Proof.** Apply Theorem 4.1 where  $H$  is a graph consisting of two vertices connected by a single edge. Take  $G'$  to be any graph on  $n/2$  nodes with  $m$  edges, and take  $\mu = 3\varepsilon m$ . Observe that a  $(1 \pm \varepsilon)$  multiplicative approximation to the number of edges in a graph distinguishes graphs with  $m$  edges from those with at least  $(1 + 3\varepsilon)m$  edges for any  $\varepsilon < 1/3$ . ◀

► **Corollary 4.3** (cf. [9, 10]). *Suppose  $\mathcal{A}$  is an algorithm that gives a  $(1 + \varepsilon)$  multiplicative approximation to the number of  $r$ -cliques in a graph using neighbor, degree, and pair queries. Specifically, for any  $\varepsilon > 0$ , on any input graph  $G = (V, E)$  with  $|V| = n$  containing  $C_r$   $r$ -cliques,  $\mathcal{A}$  outputs an estimate  $\widehat{C}_r$  satisfying  $\Pr(|\widehat{C}_r - C_r| < \varepsilon) \geq 2/3$ . Then the expected query complexity of  $\mathcal{A}$  is  $\Omega(n/(\varepsilon C_r)^{1/r})$ .*

► **Remark 4.4.** The lower bound of Corollary 4.2 is tight, as a matching upper bound is given in [18]. In the full version [12], we apply Theorem 3.3 to show that any  $(2 - \varepsilon)$  approximation to  $m$  requires  $\Omega(n^2/m)$  queries if only degree queries are allowed. This fact was observed by Feige [14], who also showed that  $O(n/\sqrt{m})$  degree queries are sufficient to obtain a 2-approximation of  $m$ .

The lower bound of Corollary 4.3 is tight for some ranges of the parameters  $n$ ,  $m$ , and  $C_r$ , but not for the entire range (see [9, 10]). In Section 4.3, we apply Theorem 3.3 (and Corollary 4.3) to obtain a tight lower bound for approximating the number of triangles in a graph ( $C_3$ ) over the entire range of parameters, thereby proving the lower bound of [9]. In the full version of this paper [12], we apply the same methodology to prove the lower bound of [10] for general  $r$ .

## 4.2 Sampling Edges

In this section, we prove a lower bound on the number of queries necessary to sample an edge in a graph  $G = (V, E)$  from an “almost-uniform” distribution  $D$  over  $E$ . The lower bound we obtain – originally proven in [13] – is tight, as a matching upper bound is proven in [13]. Here, we use “almost uniform” in the sense of total variational distance:

► **Definition 4.5.** Let  $D$  and  $D'$  be probability distributions over a finite set  $X$ . Then the **total variational distance** between  $D$  and  $D'$  is defined by

$$\text{dist}_{\text{TV}}(D, D') = \frac{1}{2} \sum_{x \in X} |D(x) - D'(x)|.$$

For  $\varepsilon > 0$ , we say that  $D$  is  **$\varepsilon$ -close to uniform** if  $\text{dist}_{\text{TV}}(D, U) \leq \varepsilon$  where  $U$  is the uniform distribution on  $X$  (i.e.,  $U(x) = 1/|X|$  for all  $x \in X$ ).

► **Theorem 4.6** (cf. [13]). *Let  $0 < \varepsilon < 1/3$ . Suppose  $\mathcal{A}$  is an algorithm that for any graph  $G = (V, E)$  on  $n$  vertices and  $m$  edges returns an edge  $e \in E$  sampled from a distribution  $D$  that is  $\varepsilon$ -close to uniform using neighbor, degree, and pair queries. Then  $\mathcal{A}$  requires  $\Omega(n/\sqrt{m})$  queries.*

**Proof.** We use the same embedding  $\mathcal{E}$  of  $\text{disj}$  described in the proof of Theorem 4.1 where  $G'$  is any graph on  $m'$  edges, and  $\ell = \sqrt{m'}$  so that  $N = n/2\sqrt{m}$ . Thus, if any  $K_j$  is a clique, the induced subgraph on  $K = K_1 \cup \dots \cup K_N$  contains at least  $m/2$  edges in  $G$  (where  $m$  is the number of edges in  $G$ ). We then take  $g : \mathcal{G}_n \rightarrow \{0, 1\}$  to be the function whose value is 0 if and only if  $K$  contains an edge. It is clear that  $(\mathcal{E}, g)$  is an embedding of  $\text{disj}$ . Thus, by Theorem 3.3 (and the proof of Theorem 4.1), any algorithm  $\mathcal{A}'$  that computes  $g$  requires  $\Omega(N) = \Omega(n/\sqrt{m})$  queries.

Let  $\mathcal{A}$  be an algorithm as in the statement of the theorem. We will show that by invoking  $\mathcal{A}$   $O(1)$  times, we can compute  $g$ . Thus, the lower bound on the number of queries for  $\mathcal{A}$  follows from the lower bound on any algorithm  $\mathcal{A}'$  computing  $g$ , as above.  $\mathcal{A}'$  works as follows: repeat  $\mathcal{A}$  7 times to get edge samples  $e_1, \dots, e_7$ . If at least one  $e_i$  satisfies  $e_i \in K \times K$ , return 0, otherwise return 1. We claim that this procedure computes  $g$  (on the range of  $\mathcal{E}$ ). To see this, suppose  $g(G) = 0$ , i.e., at least one of the  $K_i$  is a clique so that  $K \times K$  contains at least  $m/2$  edges. Thus the fraction of edges in  $K \times K$  is at least  $1/2$ , so each invocation of  $\mathcal{A}$  must return an edge  $e \in K \times K$  with probability at least  $1/2 - 1/3 = 1/6$ . Therefore, if  $g(G) = 0$ , the probability that algorithm  $\mathcal{A}'$  returns 1 (i.e., that no edge  $e \in K \times K$  is sampled) is at most  $(1 - 1/6)^7 < 1/3$ . On the other hand, if  $g(G) = 1$ , then the procedure will always return 1, as  $G$  contains no edges in  $K \times K$ . ◀

### 4.3 Counting Triangles

In this section, we prove lower lower bounds for approximately counting the number of triangles,  $C_3$  in a graph. When combined with Corollary 4.3 (with  $r = 3$ ), the main result in this section gives tight lower bounds for all ranges of the parameters  $n$ ,  $m$ , and  $C_3$ . The lower bounds (and matching upper bounds) were originally described in [9].

► **Theorem 4.7** (cf. [9]). *Let  $G$  be a graph with  $n$  vertices,  $m$  edges, and  $C_3$  triangles. Then any algorithm  $\mathcal{A}$  that computes a multiplicative approximation for  $C_3$  must perform  $\Omega\left(\min\left\{m, \frac{m^{3/2}}{C_3}\right\}\right)$  degree, neighbor, or pair queries. This lower bound holds even if  $\mathcal{A}$  is allowed to perform random edge queries (i.e.,  $\mathcal{A}$  may sample a random edge in  $G$  from a uniform distribution).*

We prove Theorem 4.7 by applying Theorem 3.3 with  $f = \text{int}_k$ , where  $N$  is the size of the instance of  $\text{int}_k$ . For any choice of the parameters  $n$ ,  $m$ , and  $C_3$ , we construct an embedding  $(\mathcal{E}, g)$  of  $\text{int}_k$  such that if  $\text{int}_k(x, y) = 1$ , then  $\mathcal{E}(x, y)$  has (roughly)  $C_3$  triangles; if  $\text{int}_k(x, y) = 0$ , then  $\mathcal{E}(x, y)$  is triangle-free.

**Proof.** Let  $n, m$ , and  $C_3$  be given. In order to simplify our presentation, we assume that  $C_3 \geq \frac{1}{2}\sqrt{m}$ .<sup>9</sup> Let  $\ell$  be a parameter (to be chosen later), and  $N = \ell^2$  the size of an instance of  $\text{int}_k$ . We identify the set  $\{0, 1\}^N$  with  $\{0, 1\}^{\ell \times \ell}$ , so that elements  $x \in \{0, 1\}^N$  are indexed by two parameters  $x = (x_{ij})$  with  $1 \leq i, j \leq \ell$ .

For  $x, y \in \{0, 1\}^N$ , we define  $G = (V, E) = \mathcal{E}(x, y)$  as follows. We partition the vertex set  $V$  into 5 sets  $A, A', B, B', S$  each of size  $\ell$ , along with an auxiliary set  $C$  of size  $n - 5\ell$ . The set  $C$  plays no role in our construction except to control the number of vertices in  $G$ . We denote  $A = \{a_1, a_2, \dots, a_\ell\}$ ,  $A' = \{a'_1, a'_2, \dots, a'_\ell\}$ , and similarly for the remaining sets in the partition. The edge set  $E$  is constructed as follows:

- For all  $a \in A, b \in B$  and  $s \in S$ , we have  $(a, s), (b, s) \in E$ .
- For all  $i, j \in [\ell]$  we have

$$\begin{cases} (a_i, b_j), (a'_j, b'_i) \in E & \text{if } x_{ij} = y_{ij} = 1 \\ (a_i, a'_j), (b_j, b'_i) \in E & \text{otherwise.} \end{cases}$$

Define the partial function  $g : \mathcal{G}_n \rightarrow \{0, 1\}$  by  $g(G) = 0$  if  $C_3(G) = 0$ , and  $g(G) = 1$  if  $C_3(G) \geq k\ell$ . In Appendix B (Claim B.1), we prove that  $(\mathcal{E}, g)$  is an embedding of  $\text{int}_k$ . To apply Theorem 3.3, we must show that each allowable query can be simulated by Alice and Bob holding  $x$  and  $y$ , respectively using  $O(1)$  bits of communication. A proof of this fact is deferred to the appendix (Claim B.2).

Since all queries can be simulated using  $O(1)$  bits of communication between Alice and Bob and  $N = \ell^2$ , Theorem 3.3 (together with communication lower bound for  $\text{int}_k$ , Corollary 2.7) implies that computing  $g$  requires  $\Omega(\ell^2/k)$  queries. For a  $k$ -intersecting instance (i.e.,  $\text{int}_k(x, y) = 1$ ), we have  $m = 4\ell^2$  and  $C_3(G) \geq k\ell$ . Thus, setting  $\ell = \frac{1}{2}\sqrt{m}$  and  $k = C_3/\ell$ , we obtain the desired result when  $2C_3/\sqrt{m} \geq 1$ . In the case where  $C_3 < \frac{1}{2}\sqrt{m}$ , we may take  $k = 1$  and modify the construction so that  $|S| = C_3$ . ◀

► **Remark 4.8.** Combined with Corollary 4.3 with  $r = 3$ , Theorem 4.7 gives a lower bound of  $\Omega\left(\min\{m, m^{3/2}/C_3\} + n/C_3^{1/3}\right)$  queries for any algorithm that obtains a multiplicative approximation to  $C_3$  in the general graph model with no random edge samples. This lower bound is tight by the matching upper bound in [9]. When random edge samples are allowed, however, the lower bound construction of Theorem 4.1 (hence Corollary 4.3) does not yield the same lower bound. To see this, note that a random edge sample will be in  $K_j$  with probability  $\varepsilon$ , so only  $O(1/\varepsilon)$  such samples are sufficient distinguish  $\text{disj}(x, y) = 1$  from 0-instances with constant probability. When random edge samples are allowed, we conjecture that the lower bound of Theorem 4.7 is tight for the entire range of  $n, m$ , and  $C_3$ .

#### 4.4 Computing Edge Connectivity

In this section, we consider the problem of estimating the edge connectivity of a graph. Recall that a graph  $G = (V, E)$  is  $k$ -(edge)-connected if at least  $k$  edges must be removed from  $G$  in order to disconnect it. Equivalently,  $G$  is  $k$ -connected if for every  $u, v \in V$ , there are at least  $k$  edge-disjoint paths between  $u$  and  $v$ . We prove the following lower bound for determining the connectivity of a graph  $G$ .

► **Theorem 4.9.** *For  $k \geq 1$ , let  $G$  be a graph with  $n$  vertices and  $m \geq 2kn$  edges. Then any algorithm  $\mathcal{A}$  that distinguishes between the case where  $G$  is  $k$ -connected and  $G$  is disconnected requires  $\Omega(m/k)$  degree, neighbor, or pair queries. This lower bound holds even if  $\mathcal{A}$  is allowed to perform random edge queries.*

<sup>9</sup> At the end of the proof we will discuss how to avoid this unnecessary assumption.

**Proof.** The proof uses a similar construction and analysis to that of Theorem 4.7. Again, we describe an embedding  $(\mathcal{E}, g)$  of  $\text{int}_k$ , for which we provide full details. The correctness of the embedding and simulation arguments are omitted, as they are nearly identical to those in the proof of Theorem 4.7.

Let  $\ell \geq 2k$  be parameter to be chosen later and  $N = \ell^2$ . Again we identify  $\{0, 1\}^N = \{(x_{ij}) \mid 1 \leq i, j \leq \ell\}$ . For  $x, y \in \{0, 1\}^N$ , the graph  $\mathcal{E}(x, y) = (V, E)$  is constructed as follows. We partition  $V$  into 5 sets,  $V = A \cup A' \cup B \cup B' \cup C$  where  $|A| = |A'| = |B| = |B'| = \ell$ , and  $|C| = n - 4\ell$ . Each  $v \in C$  is connected to  $k$  distinct vertices in  $A$  arbitrarily so that  $d(v) = k$ . We then construct edges between  $A, A', B,$  and  $B'$  according to the following rule:

$$\begin{cases} (a_i, b'_j), (b_i, a'_j) \in E & \text{if } x_{ij} = y_{ij} = 1 \\ (a_i, a'_j), (b_i, b'_j) \in E & \text{otherwise.} \end{cases} \quad (2)$$

We define the partial function  $g : \mathcal{G}_n \rightarrow \{0, 1\}$  by  $g(G) = 1$  if  $G$  is  $k$ -connected, and 0 if  $G$  is disconnected.

We claim that  $(\mathcal{E}, g)$  is an embedding of  $\text{int}_k$  () – a proof appears in Appendix C. As in the proof of Theorem 4.7, every degree, neighbor, pair, or random edge query can be simulated by Alice and Bob using at most 2 bits of communication. Therefore, by Theorem 3.3 and Corollary 2.7,  $\mathcal{A}$  requires  $\Omega(N/k) = \Omega(\ell^2/k)$  queries. The construction above satisfies  $m = 2\ell^2 + k(n - 4\ell)$  so taking  $\ell = k + \sqrt{k^2 + (m - kn)}/2 = \Theta(\sqrt{m})$  gives the desired result.  $\blacktriangleleft$

## 5 Discussion

In this paper, we presented a new technique for proving query lower bounds for graph parameter estimation problems. Here, we conclude with some open questions and suggestions for further work.

**The power of random edge queries.** In [2], Aliakbarpour et al. consider a graph query model that allows uniform random edge samples as one of its basic queries. This model is strictly stronger than the “general graph” query model that allows only degree, neighbor, and pair queries: [2] provides an upper bound for counting the number of star subgraphs in a graph that beats the lower bound described in [20] for the general graph model. In Theorems 4.7 and 4.9, our lower bounds apply to both the general graph model, as well as the stronger model with uniform random edge samples. In the constructions described in the proofs of Theorems 4.7 and 4.9, edge samples do not afford more computational power, essentially because the degree sequence of the constructions is fixed. In particular, each random edge sample can be simulated by sampling a vertex with probability proportional to its (known) degree, and using a neighbor query to sample a random incident edge. Indeed, sampling edges from a uniform distribution is equivalent to sampling vertices with probability proportional to their degrees.

The construction used for lower bound of Theorem 4.1 cannot give a lower bound better than  $\Omega(\min\{m/\mu^{2/k}, n/\mu^{1/k}\})$ , as a clique  $K_j$  in that construction contains  $\mu^{2/k}$  edges. In particular, the lower bound for estimating  $m$  (Corollary 4.2) becomes only  $\Omega(1/\varepsilon)$  if random edge samples are used. An upper bound of  $O(n^{1/3})$  for estimating  $m$  with random edge (and vertex) samples is implied by the algorithm of Motwani et al. [27]. The authors also prove a lower bound of  $\Omega(n^{1/3})$ , although the construction only holds for  $m = O(n^{2/3})$ . We believe it is an interesting problem to characterize the complexity of estimating  $m$  in the general graph model with random edge samples the over the full range of  $m$ .

In general, we would like to better understand the power uniform random edge samples in the general graph model. We conjecture that the lower bound of Theorem 4.7 is tight over the entire range of the parameters. The algorithm of Eden et al. [9] proves that the lower bound is tight for  $\min\{m, m^{3/2}/C_3\} = \Omega(n/C_3^{1/3})$  even without edge samples. Thus, edge samples may only help in the regime where  $n/C_3^{1/3} = \omega(\min\{m, m^{3/2}/C_3\})$ .

► **Question 5.1.** For what graph parameter estimation problems do random edge samples help?

**Property testing lower bounds.** The graph query access models we consider were initially proposed in the context of property testing [16, 17, 28, 25]. In graph property testing with the general graph model [25], the goal is to distinguish graphs that satisfy some property  $P$  from those that are far from satisfying  $P$  in the sense that an  $\varepsilon$ -fraction of edges of the graph must be modified in order to the graph to satisfy  $P$ . In this model, our constructions imply property testing lower bounds, at least for some range of  $m$ . For example, we state a consequence of Theorem 4.1 for testing the property of triangle-freeness (i.e., that  $C_3(G) = 0$ ).

► **Corollary 5.2.** *Any property testing algorithm for triangle-freeness in the general graph model requires  $\Omega(n/\sqrt{\varepsilon m})$  queries.*

Corollary 5.2 follows from the construction in Theorem 4.1 by taking  $G'$  to be any triangle-free graph on  $n/2$  vertices and  $m$  edges, and  $\ell = \sqrt{\varepsilon m}$ . This way each (potential) clique  $K_j$  on  $\ell$  vertices will contain roughly  $\varepsilon m$  edges. Further, by Turán's theorem [33, 1], at least (roughly)  $\varepsilon/3m$  edges must be removed from  $G$  (in particular from  $K_j$ ) in order to make  $G$  triangle-free in the case where  $K_j$  is a clique. The lower bound of Corollary 5.2 matches the known lower bound due to Alon et al. [3] in the regime where the average degree  $d = 2m/n$  satisfies  $d = O(n^{1/3})$ . In the range  $d = \omega(n^{1/3})$ , the lower bounds in [3] are strictly stronger.<sup>10</sup>

In the dense graph model [16], only pair queries are allowed, but the distances between graphs are normalized by  $n^2$  (rather than  $|E|$  as in the general graph model). Thus every graph with  $m = o(n^2)$  is  $\varepsilon$ -close to the graph with no edges. In this case, the types of embeddings we present in this paper cannot yield lower bounds that are better than  $\Omega(1/\varepsilon)$ . Specifically, in all embeddings of disjointness we consider, each edge in  $\mathcal{E}(x, y)$  depends on a single bit of  $x, y \in \{0, 1\}^N$ . However, the value of  $\text{disj}(x, y)$  can vary by changing a single bit of  $x$  or  $y$ . In order for Theorem 3.3 to give property testing lower bounds via an embedding of  $\text{disj}$ , changing a single bit of  $x$  or  $y$  must change  $\varepsilon n^2$  edges in  $\mathcal{E}(x, y)$ . Thus, to obtain stronger lower bounds (e.g., lower bounds that grow as a function of  $n$ ), either different communication primitives must be considered, or the embedding  $\mathcal{E}$  must be more complicated (using some nontrivial encoding of  $x$  and  $y$ ).

► **Question 5.3.** Can Theorem 3.3 be applied to obtain any nontrivial (i.e.,  $\omega(1/\varepsilon)$ ) lower bound for any “natural” graph problem in the dense graph property testing model?

<sup>10</sup>Alon et al. [3] prove that a lower bound of  $\Omega(n^{1/3})$  holds even for all  $d = O(n^{1-\nu(n)})$  for some function  $\nu(n) = o(1)$ . This is in contrast for the case  $d = \Omega(n)$ , where  $O(f(\varepsilon))$  queries are sufficient for some function  $f$ .

## References

- 1 Martin Aigner and Günter M Ziegler. *Proofs from the Book*. Springer-Verlag Berlin Heidelberg, 4th edition, 2010. doi:10.1007/978-3-642-00856-6.
- 2 Maryam Aliakbarpour, Amartya Shankha Biswas, Themistoklis Gouleakis, John Peebles, Ronitt Rubinfeld, and Anak Yodpinyanee. Sublinear-time algorithms for counting star subgraphs with applications to join selectivity estimation. *CoRR*, abs/1601.04233, 2016. arXiv:1601.04233.
- 3 Noga Alon, Tali Kaufman, Michael Krivelevich, and Dana Ron. Testing triangle-freeness in general graphs. *SIAM Journal on Discrete Mathematics*, 22(2):786–819, 2008.
- 4 Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pages 20–29. ACM, 1996.
- 5 Ziv Bar-Yossef, T.S. Jayram, Ravi Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. *Journal of Computer and System Sciences*, 68(4):702–732, 2004. Special Issue on FOCS 2002. doi:10.1016/j.jcss.2003.11.006.
- 6 Suman K. Bera and Amit Chakrabarti. Towards Tighter Space Bounds for Counting Triangles and Other Substructures in Graph Streams. In Heribert Vollmer and Brigitte Vallée, editors, *34th Symposium on Theoretical Aspects of Computer Science (STACS 2017)*, volume 66 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 11:1–11:14, Dagstuhl, Germany, 2017. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. doi:10.4230/LIPIcs.STACS.2017.11.
- 7 Eric Blais, Joshua Brody, and Kevin Matulef. Property testing lower bounds via communication complexity. *computational complexity*, 21(2):311–358, Jun 2012. doi:10.1007/s00037-012-0040-x.
- 8 Shahar Dobzinski, Noam Nisan, and Michael Schapira. Approximation algorithms for combinatorial auctions with complement-free bidders. In *Proceedings of the Thirty-seventh Annual ACM Symposium on Theory of Computing*, STOC '05, pages 610–618, New York, NY, USA, 2005. ACM. doi:10.1145/1060590.1060681.
- 9 T. Eden, A. Levi, D. Ron, and C. Seshadhri. Approximately counting triangles in sublinear time. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 614–633, Oct 2015. doi:10.1109/FOCS.2015.44.
- 10 Talya Eden, Dana Ron, and C. Seshadhri. On approximating the number of  $k$ -cliques in sublinear time. *CoRR*, abs/1707.04858, 2017. arXiv:1707.04858.
- 11 Talya Eden, Dana Ron, and C. Seshadhri. Sublinear time estimation of degree distribution moments: The degeneracy connection. In *44th International Colloquium on Automata, Languages, and Programming, ICALP 2017, July 10-14, 2017, Warsaw, Poland*, pages 7:1–7:13, 2017. Full version available at <https://arxiv.org/abs/1604.03661>. doi:10.4230/LIPIcs.ICALP.2017.7.
- 12 Talya Eden and Will Rosenbaum. Lower bounds for approximating graph parameters via communication complexity. Full version of this paper, 2017. URL: <https://arxiv.org/abs/1709.04262>.
- 13 Talya Eden and Will Rosenbaum. On Sampling Edges Almost Uniformly. In Raimund Seidel, editor, *1st Symposium on Simplicity in Algorithms (SOSA 2018)*, volume 61 of *Open-Access Series in Informatics (OASICS)*, pages 7:1–7:9, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. Full version available at <https://arxiv.org/abs/1706.09748>. doi:10.4230/OASICS.SOSA.2018.7.
- 14 Uri Feige. On sums of independent random variables with unbounded variance and estimating the average degree in a graph. *SIAM J. Comput.*, 35(4):964–984, 2006. doi:10.1137/S0097539704447304.



- 15 Oded Goldreich. On the communication complexity methodology for proving lower bounds on the query complexity of property testing. *Electronic Colloquium on Computational Complexity (ECCC)*, 20:73, 2013. URL: <http://eccc.hpi-web.de/report/2013/073>.
- 16 Oded Goldreich, Shari Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. *Journal of the ACM (JACM)*, 45(4):653–750, 1998.
- 17 Oded Goldreich and Dana Ron. Property testing in bounded degree graphs. *Algorithmica*, pages 302–343, 2002.
- 18 Oded Goldreich and Dana Ron. Approximating average parameters of graphs. *Random Structures & Algorithms*, 32(4):473–493, 2008. doi:10.1002/rsa.20203.
- 19 Yannai A. Gonczarowski, Noam Nisan, Rafail Ostrovsky, and Will Rosenbaum. A stable marriage requires communication. In *Proceedings of the Twenty-sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '15, pages 1003–1017, Philadelphia, PA, USA, 2015. Society for Industrial and Applied Mathematics. URL: <http://dl.acm.org/citation.cfm?id=2722129.2722197>.
- 20 Mira Gonen, Dana Ron, and Yuval Shavitt. Counting stars and other small subgraphs in sublinear-time. *SIAM Journal on Discrete Mathematics*, 25(3):1365–1411, 2011.
- 21 J. Hromkovič. *Communication Complexity and Parallel Computing*. Texts in Theoretical Computer Science. An EATCS Series. Springer Berlin Heidelberg, 2013.
- 22 Bala Kalyanasundaram and Georg Schintger. The probabilistic communication complexity of set intersection. *SIAM Journal on Discrete Mathematics*, 5(4):545–557, 1992.
- 23 Bala Kalyanasundaram and Georg Schintger. *Communication Complexity and Lower Bounds for Sequential Computation*, pages 253–268. Vieweg+Teubner Verlag, Wiesbaden, 1992. doi:10.1007/978-3-322-95233-2\_15.
- 24 Mauricio Karchmer and Avi Wigderson. Monotone circuits for connectivity require super-logarithmic depth. *SIAM Journal on Discrete Mathematics*, 3(2):255–265, 1990. doi:10.1137/0403021.
- 25 Tali Kaufman, Michael Krivelevich, and Dana Ron. Tight bounds for testing bipartiteness in general graphs. *SIAM Journal on computing*, 33(6):1441–1483, 2004.
- 26 E. Kushilevitz and N. Nisan. *Communication Complexity*. Cambridge University Press, 2006.
- 27 Rajeev Motwani, Rina Panigrahy, and Ying Xu. Estimating sum by weighted sampling. In *ICALP*, volume 4596, pages 53–64. Springer, 2007.
- 28 Michal Parnas and Dana Ron. Testing the diameter of graphs. *Random Structures & Algorithms*, 20(2):165–183, 2002. doi:10.1002/rsa.10013.
- 29 Ramamohan Paturi and Janos Simon. Probabilistic communication complexity. *Journal of Computer and System Sciences*, 33(1):106–123, 1986.
- 30 Alexander A. Razborov. On the distributional complexity of disjointness. *Theoretical Computer Science*, 106(2):385–390, 1992.
- 31 Atish Das Sarma, Stephan Holzer, Liah Kor, Amos Korman, Danupon Nanongkai, Gopal Pandurangan, David Peleg, and Roger Wattenhofer. Distributed verification and hardness of distributed approximation. *SIAM Journal on Computing*, 41(5):1235–1265, 2012. doi:10.1137/11085178X.
- 32 Jacob Steinhardt, Gregory Valiant, and Stefan Wager. Memory, communication, and statistical queries. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1490–1516, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR. URL: <http://proceedings.mlr.press/v49/steinhardt16.html>.
- 33 Paul Turán. On an extremal problem in graph theory. *Matematikai és Fizikai Lapok*, 48:436–452, 1941.

- 34 Andrew Chi-Chih Yao. Some complexity questions related to distributive computing (preliminary report). In *Proceedings of the Eleventh Annual ACM Symposium on Theory of Computing*, STOC '79, pages 209–213, New York, NY, USA, 1979. ACM. doi:10.1145/800135.804414.

## A $k$ -Intersection Lower Bound

**Proof of Corollary 2.7.** The argument is by simulation. Specifically, we will show that any efficient protocol for  $\text{int}_k$  yields an efficient protocol for  $\text{disj}$ . Suppose  $\Pi$  is a protocol for the promise problem of the corollary with  $\text{cost}(\Pi) = B$ . For  $x, y \in \{0, 1\}^{N/k}$ , let  $x^k, y^k \in \{0, 1\}^N$  denote the concatenation of  $x$  and  $y$  (respectively) repeated  $k$  times. Observe that if  $x$  and  $y$  satisfy the unique intersection promise, then  $x^k$  and  $y^k$  satisfy the  $k$ -intersection promise. Further,  $\text{int}_k(x^k, y^k) = 0$  if and only if  $\text{disj}(x, y) = 1$ . Since  $\Pi$  computes  $\text{int}_k$  for all  $x', y' \in \{0, 1\}^N$  satisfying the  $k$ -intersection promise,  $\Pi(x^k, y^k)$  computes  $\neg \text{disj}$  on input  $x, y$ . Therefore, by Theorem 2.5,  $\text{cost}(\Pi) = \Omega(N/k)$ , which gives the desired result. ◀

## B Theorem 4.7 Details

In this appendix, we provide details to complete the proof of Theorem 4.7.

► **Claim B.1.** Let  $\mathcal{E}$  be defined as in Section 4.3, and define the partial function  $g : \mathcal{G} \rightarrow \{0, 1\}$  by

$$g(G) = \begin{cases} 0 & \text{if } C_3(G) = 0 \\ 1 & \text{if } C_3(G) \geq k\ell. \end{cases}$$

Then  $(\mathcal{E}, g)$  is an embedding of  $\text{int}_k$ .

**Proof.** To see the claim is true, first consider the case where  $\text{int}_k(x, y) = 0$  – i.e.,  $x$  and  $y$  are disjoint. In this case, all edges in  $\mathcal{E}(x, y)$  are between  $A$  and  $S$ ,  $A$  and  $A'$ ,  $B$  and  $S$ , or  $B$  and  $B'$ . Therefore there are no edges between  $S \cup A \cup B$  and  $A' \cup B' \cup C$ , implying that  $G$  is bipartite, hence triangle-free. On the other hand, if  $\text{int}_k(x, y) = 1$ , then for each of the (at least  $k$ ) pairs  $(i, j) \in [\ell]^2$  satisfying  $x_{ij} = y_{ij} = 1$ , we have  $(a_i, b_j) \in E$ . Therefore, for each  $s \in S$ , the edges  $(a_i, b_j), (b_j, s), (s, a_i) \in E$  form a triangle. Since  $|S| = \ell$ , this implies that  $\mathcal{E}(x, y)$  contains at least  $k\ell$  triangles. ◀

► **Claim B.2.** For the embedding  $(\mathcal{E}, g)$  described in Section 4.3, any degree, neighbor, or pair query, and a random edge sampling can be simulated using  $O(1)$  bits of communication between Alice and Bob.

**Proof.** We consider each allowable query separately. For any  $u, v \in V$ , the queries can be simulated as follows.

**Degree query.**  $d(u)$  is independent of  $x$  and  $y$ : if  $u \in S \cup A \cup B$ , then  $d(u) = 2\ell$ ; if  $u \in A' \cup B'$ , then  $d(u) = \ell$ ; if  $u \in C$ , then  $d(u) = 0$ . Thus Alice and Bob can simulate any degree query without communication.

**Neighbor query.** For  $a_i \in A$ , we label  $a_i$ 's incident edges so that  $a_i$ 's  $j^{\text{th}}$  neighbor is either  $b_j$  (if  $x_{ij} = y_{ij} = 1$ ) or  $a'_j$  otherwise. Edges incident to  $A'$ ,  $B$ , and  $B'$  are labeled similarly. Thus, Alice and Bob can answer queries of the form  $\text{nbr}_i(u)$  for  $v \in A \cup A' \cup B \cup B'$  with  $j \leq \ell$  by exchanging  $x_{ij}$  and  $y_{ij}$  using 2 bits of communication. All other neighbor queries can be answered without communication.

**Pair query.** Alice and Bob can answer pair queries of the form  $\text{pair}(a_i, a'_j)$ ,  $\text{pair}(a_i, b_j)$ , and  $\text{pair}(b_j, b'_i)$  by exchanging  $x_{ij}$  and  $y_{ij}$ . All other queries can be answered without communication. Again the communication cost is 2 bits.

**Uniform edge sample.** Alice and Bob can sample a random edge from a uniform distribution using their shared public randomness and the fact that each node in  $\mathcal{E}(x, y)$  has the same degree independent of  $x$  and  $y$ . To this end, Alice and Bob sample  $e = (u, v)$  by first sampling a vertex  $v \in V$  where each node is chosen with probability proportional to its degree,  $d(v)$ . Alice and Bob then choose a random number  $i \in [d(v)]$  uniformly at random, and sample the edge  $e = (v, u)$  where  $u = \text{nbr}_i(v)$  at a communication cost of (at most) 2 bits. Note that  $e = (v, u)$  is sampled with probability

$$\Pr(e = (v, u) \text{ is sampled}) = \frac{d(v)}{\sum_{w \in V} d(w)} \cdot \frac{1}{d(v)} + \frac{d(u)}{\sum_{w \in V} d(w)} \frac{1}{d(u)} = \frac{1}{m},$$

so that edges are indeed sampled according to a uniform distribution. ◀

## C Theorem 4.9 Details

► **Claim C.1.** *Let  $\mathcal{E}$  and  $g$  be as described in Section 4.4. Then  $(\mathcal{E}, g)$  is an embedding of  $\text{int}_k$  (where we assume the promise that  $\sum_{i,j} x_{ij}y_{ij} \in \{0, k\}$ ).*

**Proof.** In the case where  $\text{int}_k(x, y) = 0$ , there are no edges between  $A \cup A' \cup C$  and  $B \cup B'$ , hence  $\mathcal{E}(x, y)$  is disconnected. If  $\text{int}_k(x, y) = 1$ , we must show that  $\mathcal{E}(x, y)$  is  $k$ -connected. We will show that there are at least  $k$  edge disjoint paths between any pair of vertices in  $\mathcal{E}(x, y)$ . We consider the following cases separately.

**Case 1:**  $u, v \in A$  (or symmetrically,  $u, v \in A', B$ , or  $B'$ ). From the definition of  $\mathcal{E}$  (Equation (2)) and the promise for  $\text{int}_k$ , there are at most  $k$  pairs  $(a_i, a'_j) \in A \times A'$  that are *not* contained in  $E$ . Since  $\ell \geq 2k$ , this implies that  $u, v \in A$  have at least  $\ell - k \geq k$  common neighbors in  $A'$ . In particular, there are at least this many edge disjoint paths (of length 2) between  $u$  and  $v$ .

**Case 2:**  $u \in A, v \in A'$  (or symmetrically,  $u \in B, v \in B'$ ). As before,  $v$  has at least  $k$  distinct neighbors,  $u_1, \dots, u_k \in A$ . Further, by the analysis in Case 1, each  $u_i$  has at least  $k$  common neighbors with  $u$ . Therefore there exists a matching  $(u_1, v_1), \dots, (u_k, v_k) \in E$ .<sup>11</sup> The paths  $(u, v_i), (v_i, u_i), (u_i, v)$  for  $i = 1, \dots, k$  are then edge disjoint. (Note that it may be the case that  $u_i = u$ , in which case we take  $(u, v)$  to be one of the matching edges and take this edge to be the corresponding path between  $u$  and  $v$ .)

**Case 3:**  $u \in A, v \in B'$  (or symmetrically  $u \in A', v \in B$ ). Let  $(u_1, v_1), \dots, (u_k, v_k) \in A' \times B$  be the  $k$  edges between  $A'$  and  $B$ . Take  $U = \{u_1, \dots, u_k\}$  (respectively,  $V = \{v_1, \dots, v_k\}$ ) to be the multiset of endpoints of the edges between  $A'$  and  $B$  in  $A'$  (respectively  $B$ ). It suffices to show that there are edge-disjoint paths from  $u$  to each  $u_i$  (with multiplicity) – the analogous result for  $v$  and  $v_i$  is identical. Let  $u'_1, u'_2, \dots, u'_k$  be distinct neighbors of  $u$ . Then, as in Case 1, each  $u_i$  and  $u'_i$  have at least  $k$  common neighbors. By choosing one such common neighbor,  $u''_i$  for each  $i$  (greedily) we can form  $k$  edge disjoint paths  $(u, u'_i), (u'_i, u''_i), (u''_i, u_i)$ .

**Case 4:**  $u \in A, v \in B$  (or symmetrically,  $u \in A', v \in B'$ ). Let  $(u_1, v_1), \dots, (u_k, v_k)$  be as in Case 3. As in Case 3, there are  $k$  edge-disjoint paths in  $A \cup A'$  from  $u$  to the  $u_i$ . Further, there are  $k$  edge-disjoint paths from  $v$  to the  $v_j$  as in Case 2.

<sup>11</sup> Such a matching can be found by greedily choosing common neighbors of  $u$  and  $u_1, u_2$ , etc.

## 11:18 Lower Bounds for Approximating Graph Parameters via Communication Complexity

**Case 5:**  $u \in C$ . Let  $u_1, \dots, u_k$  be the neighbors of  $u$  in  $A$ . It suffices to show that there are edge disjoint paths from each  $u_i$  to  $v$ . The cases  $v \in A, A', B, B'$  are analogous to arguments in Cases 1–4. If  $v \in C$ , let  $v_1, \dots, v_k$  be the neighbors of  $v$  in  $A$ . Since each pair  $u_i, v_i$  share  $k$  common neighbors in  $A'$ , we can assign a unique neighbor  $w_i$  to each such pair so that  $(u, u_i), (u_i, w_i), (w_i, v_i), (v_i, v)$  for  $i = 1, \dots, k$  are edge disjoint paths. ◀