Algorithmic Aspects of Property Testing in the Dense Graphs Model

Oded Goldreich*
Department of Computer Science
Weizmann Institute of Science
Rehovot, ISRAEL.
oded.goldreich@weizmann.ac.il

Dana Ron†
Department of EE–Systems
Tel-Aviv University
Ramat-Aviv, ISRAEL.
danar@eng.tau.ac.il

April 11, 2008

Abstract

In this paper we consider two refined questions regarding the query complexity of testing graph properties in the adjacency matrix model. The first question refers to the relation between adaptive and non-adaptive testers, whereas the second question refers to testability within complexity that is inversely proportional to the proximity parameter, denoted ϵ. The study of these questions reveals the importance of algorithmic design (also) in this model. The highlights of our study are:

- A gap between the complexity of adaptive and non-adaptive testers. Specifically, there exists a (natural) graph property that can be tested using $\widetilde{O}(\epsilon^{-1})$ adaptive queries, but cannot be tested using $o(\epsilon^{-3/2})$ non-adaptive queries.
- In contrast, there exist natural graph properties that can be tested using $\widetilde{O}(\epsilon^{-1})$ non-adaptive queries, whereas $\Omega(\epsilon^{-1})$ queries are required even in the adaptive case.

We mention that the properties used in the foregoing conflicting results have a similar flavor, although they are of course different.

Keywords: Property Testing, Adaptivity vs Non-adaptivity, Graph Properties,

*Partially supported by the Israel Science Foundation (grant No. 460/05).
†Partially supported by the Israel Science Foundation (grant No. 89/05).

Contents

1 Introduction

In the last decade, the area of property testing has attracted much attention (see the surveys of [F01, R01], which are already out-of-date). Loosely speaking, property testing typically refers to sub-linear time probabilistic algorithms for deciding whether a given object has a predetermined property or is far from any object having this property. Such algorithms, called testers, obtain bits of the object by making adequate queries, which means that the object is seen as a function and the testers get oracle access to this function (and thus may be expected to work in time that is sub-linear in the length of the description of this object).

Much of the aforementioned work (see, e.g., [GGR, AFKS, AFNS]) was devoted to the study of testing graph properties in the adjacency matrix model, which is also the setting of the current work. In this model, introduced in [GGR], graphs are viewed as (symmetric) Boolean functions over a domain consisting of all possible vertex-pairs (i.e., an N-vertex graph $G = ([N], E)$ is represented by the function $g : [N]\times[N] \to \{0, 1\}$ such that $\{u, v\} \in E$ if and only if $g(u, v) = 1$). Consequently, an N-vertex graph represented by the function $g : [N] \times [N] \to \{0, 1\}$ is said to be ϵ-far from some predetermined graph property if at least $\epsilon \cdot N^2$ entries of g must be modified in order to yield a representation of a graph that has this property. We refer to ϵ as the proximity parameter, and the complexity of testing is stated in terms of ϵ and the number of vertices in the graph (i.e., N).

Interestingly, many natural graph properties can be tested within query complexity that depends only on the proximity parameter; see [GGR], which presents testers with query complexity $\text{poly}(1/\epsilon)$, and [AFNS], which characterizes the class of properties that are testable within query complexity that depends only on the proximity parameter (where this dependence may be an arbitrary function of ϵ). However, a common phenomenon in all the aforementioned works is that they utilize quite naive algorithms and their focus is on the (often quite sophisticated) analysis of these algorithms. This phenomenon is no coincidence: As shown in [AFKS, GT], when ignoring a quadratic blow-up in the query complexity, property testing (in this model) reduces to sheer combinatorics. Specifically, without loss of generality, the tester may just inspect a random induced subgraph (of adequate size) of the input graph.

In this paper we demonstrate that a more refined study of property testing (in this model) reveals the importance of algorithmic design (also in this model). This is demonstrated both by studying the advantage of adaptive testers over non-adaptive ones as well as by studying the class of properties that can be tested within complexity that is inversely proportional to the proximity parameter.

1.1 Two Related Studies

Let us start by reviewing the two related studies conducted in the current work.

Adaptivity vs Non-adaptivity. A tester is called non-adaptive if it determines all its queries independently of the answers obtained for previous queries, and otherwise it is called adaptive. Indeed, by [AFKS, GT], the benefit of adaptivity (or, equivalently, the cost of non-adaptivity) is rather small: Specifically, any (possibly adaptive) tester (for any graph property) of query complexity $q(N, \epsilon)$ can be transformed into a non-adaptive tester of query complexity $O(q(N, \epsilon)^2)$. But is this quadratic gap an artifact of the known proofs (of [AFKS, GT]) or does it reflect something inherent?

A recent work by [GnRn] suggests that the latter case may hold: For every $\epsilon > 0$, they showed that the set of N-vertex bipartite graphs of maximum degree $O(\epsilon N)$ is ϵ-testable (i.e., testable with respect to proximity parameter ϵ) by $\widetilde{O}(\epsilon^{-3/2})$ queries, while (by [BT]) a non-adaptive tester for this set must use $\Omega(\epsilon^{-2})$ queries. Thus, there exists a case where non-adaptivity requires increasing the query complexity; specifically, for any $c < 4/3$, the query complexity of the non-adaptive tester is greater than a c-power of the query complexity of the adaptive tester (i.e., $\widetilde{O}(\epsilon^{-3/2})^c = o(\epsilon^{-2})$). We stress that the result of [GnRn] does not refer to property testing in the "proper" sense; that is, the complexity is not analyzed with respect to a varying value of the proximity parameter, while the property itself is fixed. It is rather the case that, for every value of the proximity parameter, a different property (which depends on this parameter) is considered and the (upper- and lower-) bounds refer to this combination (of a property tailored for a fixed value of the proximity parameter). Thus, *the work of [GnRn] leaves open the question of whether there exists a single graph property such that adaptivity is beneficial for any value of the proximity parameter* (as long as $\epsilon > N^{-\Omega(1)}$). That is, the question is whether adaptivity is beneficial for the standard asymptotic-complexity formulation of property testing.

Complexity inversely proportional to the proximity parameter. As shown in [GGR], many natural graph properties can be tested within query complexity that is polynomial in the reciprocal of the proximity parameter (and independent of the size of the graph). We ask whether a linear complexity is possible at all, and if so which properties can be tested within query complexity that is linear (or almost linear) in the reciprocal of the proximity parameter.[1]

The first question is easy to answer. Avoiding trivial properties, we note that the property of being a clique (equiv., an independent set) can be tested by $O(1/\epsilon)$ queries, even when these questions are non-adaptive (e.g., make $O(1/\epsilon)$ random queries and accept if and only if all return 1). Still, we ask whether "more interesting"[2] graph theoretical properties can also be tested within similar complexity (either only adaptively or also non-adaptively).

1.2 Our Results

We address the foregoing questions by studying a sequence of natural graph properties (defined formally in Section 2.2). The first property in the sequence, called clique collection and denoted $\mathcal{CC}$, is the set of graphs such that each graph consists of a collection of isolated cliques. Such graphs model situations in which a set of possibly related elements can be partitioned into "perfect clusters" (i.e., two elements are in the same cluster if and only if they are related). For this property (i.e., $\mathcal{CC}$), we prove a gap between adaptive and non-adaptive query complexity, where the adaptive query complexity is almost linear in the reciprocal of the proximity parameter. That is:

Theorem 1.1 (the query complexity of clique collection):

1. *There exists an adaptive tester of query complexity $\widetilde{O}(\epsilon^{-1})$ for $\mathcal{CC}$. Furthermore, this tester runs in time $\widetilde{O}(\epsilon^{-1})$.*[3]
2. *Any non-adaptive tester for $\mathcal{CC}$ must have query complexity $\Omega(\epsilon^{-4/3})$.*

[1] Note that $\Omega(1/\epsilon)$ queries are required for testing any of the graph properties considered in the current work; for a more general statement see the beginning of Section 6.

[2] A more articulated reservation towards the foregoing properties may refer to the fact that these graph properties contain a single N-vertex graph (per each N) and are represented by monochromatic functions.

[3] We refer to a model in which elementary operations regarding pairs of vertices are charged at unit cost.

3. *There exists a non-adaptive tester of query complexity $O(\epsilon^{-4/3})$ for $\mathcal{CC}$. Furthermore, this tester runs in time $O(\epsilon^{-4/3})$.*

Note that the complexity gap (between Parts 1 and 2) of Theorem 1.1 matches the gap established by [GnRn] (for "non-proper" testing). A larger gap is established for a property of graphs, called bi-clique collection and denoted $\mathcal{BCC}$, where a graph is in $\mathcal{BCC}$ if it consists of a collection of isolated bi-cliques (i.e., complete bipartite graphs).

Theorem 1.2 (the query complexity of bi-clique collection):

1. *There exists an adaptive tester of query complexity $\widetilde{O}(\epsilon^{-1})$ for $\mathcal{BCC}$. Furthermore, this tester runs in time $\widetilde{O}(\epsilon^{-1})$.*
2. *Any non-adaptive tester for $\mathcal{BCC}$ must have query complexity $\Omega(\epsilon^{-3/2})$. Furthermore, this holds even if the input graph is promised to be bipartite.*

The furthermore clause (in Part 2 of Theorem 1.2) holds also for the model studied in [AFN], where the bi-partition of the graph is given. We note that bi-cliques may be viewed as the bipartite analogues of cliques (w.r.t general graphs), and indeed they arise naturally in applications that are modeled by bipartite graphs.

Theorem 1.2 asserts that the gap between the query complexity of adaptive and non-adaptive testers may be a power of $1.5 - o(1)$. Recall that the results of [AFKS, GT] assert that the gap may not be larger than quadratic. We conjecture that this upper-bound can be matched.

Conjecture 1.3 (an almost-quadratic complexity gap): *For every positive integer $t \geq 5$, there exists a graph property Π for which the following holds:*

1. *There exists an adaptive tester of query complexity $\widetilde{O}(\epsilon^{-1})$ for Π. Furthermore, this tester runs in time $\widetilde{O}(\epsilon^{-1})$.*
2. *Any non-adaptive tester for Π must have query complexity $\Omega(\epsilon^{-2+(2/t)})$.*

Furthermore, Π consists of graphs that are each a collection of "super-cycles" of length t, where a super-cycle is a set of t independent sets arranged on a cycle such that each pair of adjacent independent sets is connected by a complete bipartite graph.

We were able to prove Part 2 of Conjecture 1.3, but failed to provide a full analysis of the algorithm intended for Part 1. We comment that we can prove a *promise problem version of Conjecture 1.3*; specifically, this promise problem (stated in Theorem 5.5) refers to inputs promised to reside in a set $\Pi' \supset \Pi$ and the tester is required to distinguish graphs in Π from graphs that are ϵ-far from Π.

In contrast to the foregoing results that aim at identifying properties with a substantial gap between the query complexity of adaptive versus non-adaptive testing, we also study cases in which no such gap exists. Since query complexity that is linear in the reciprocal of the proximity parameter is minimal for many natural properties (and, in fact, for any property that is "non-trivial for testing"), we focus on non-adaptive testers that (approximately) meet this bound. Among the results obtained in this direction, we highlight the following one.

Theorem 1.4 (the query complexity of collections of $O(1)$ cliques): *For every positive integer c, there exists a non-adaptive tester of query complexity $\widetilde{O}(\epsilon^{-1})$ for the set of graphs such that each graph consists of a collection of up to c cliques. Furthermore, this tester runs in time $\widetilde{O}(\epsilon^{-1})$.*

Discussion. The foregoing results demonstrate that a finer look at property testing of graphs in the adjacency matrix model reveals the role of algorithm design in this model. In particular, in some cases (see, e.g., Theorems 1.1 and 1.2), carefully designed adaptive algorithms outperform any non-adaptive algorithm. Indeed, this conclusion stands in contrast to [GT, Thm. 2], which suggests that a less fine view (which ignores polynomial blow-ups)[4] deems algorithm design irrelevant to this model. We also note that, in some cases (see, e.g., Theorem 1.4 and Part 3 of Theorem 1.1), carefully designed non-adaptive algorithms outperform straightforward ones.

A different perspective on this work is as a study of the relation between adaptive and non-adaptive queries. Needless to say, this fundamental relation was studied in a variety of models, and the current work studies it in a specific natural model (i.e., of property testing in the adjacency matrix representation).[5] Our results demonstrate that, in this model, the relation between the adaptive and non-adaptive query-complexities is not fixed, but rather varies with the computational problem at hand. In some cases (e.g., Theorem 1.4) the complexities are essentially equal (indeed, as in the case of sampling [CEG]). In other cases (e.g., Theorem 1.1), these complexities are related by a fixed power (e.g., $4/3$) that is strictly between 1 and 2. And, yet, in other cases (e.g., Theorem 5.5) the non-adaptive complexity is quadratic in the adaptive complexity, which is the maximum gap possible (by [AFKS, GT]). We conjecture that, for any $t \geq 3$, there exists a property for which the aforementioned complexities are related by a power of $2 - (2/t)$.

1.3 Open Problems

In addition to the resolution of Conjecture 1.3, our study raises many other open problems; the most evident ones are listed next.

1. What is the non-adaptive query complexity of $\mathcal{BCC}$? Note that Theorem 1.2 only establishes a lower-bound of $\Omega(\epsilon^{-3/2})$. We conjecture that an efficient non-adaptive algorithm of query complexity $\widetilde{O}(\epsilon^{-3/2})$ can be devised.

2. For which constants $c \in [1,2]$ does there exist a property that has adaptive query complexity of $q(\epsilon)$ and non-adaptive query complexity of $\widetilde{\Theta}(q(\epsilon)^c)$? Note that Theorem 1.1 shows that $4/3$ is such a constant, and the same holds for the constant 1 (see, e.g., Theorem 1.4). We conjecture that, for any $t \geq 2$, it holds that the constant $2 - (2/t)$ also satisfies the foregoing requirement. It may be the case that these constants are the only ones that satisfy this requirement.

3. Characterize the class of graph properties for which the query complexity of non-adaptive testers is almost linear in the query complexity of adaptive testers.

4. Characterize the class of graph properties for which the query complexity of non-adaptive testers is almost quadratic in the query complexity of adaptive testers.

[4]Recall that [GT, Thm. 2] asserts that canonical testers, which merely select a random subset of vertices and rule according to the induced subgraph, have query-complexity that is at most quadratic in the query-complexity of the best tester. We note that [GT, Thm. 2] also ignores the time-complexity of the testers.

[5]We mention that this relation has also been studied in the context of property testing (and in a variety of different settings). Specifically, in the setting of testing the satisfiability of linear constraints, it was shown that adaptivity offers absolutely no gain [BHR]. A similar result holds for testing monotonicity of Boolean functions [F04]. In contrast, an exponential gap between the adaptive and non-adaptive complexities may exist in the context of testing other properties of Boolean functions [F04]. Lastly, we mention that an even more dramatic gap exists in the setting of testing graph properties in the bounded-degree model (of [GR02]); see [RaSm].

5. Characterize the class of graph properties for which the query complexity of adaptive (resp., non-adaptive) testers is almost linear in the reciprocal of the proximity parameter.

Finally, we recall the well-known open problem (partially addressed in [AS]) of providing a characterization of the class of graph properties that are testable within query complexity that is polynomial in the reciprocal of the proximity parameter.

1.4 Organization

Section 2 contains a review of the basic notions underlying this work as well as formal definitions of the graph properties that we study. In Section 3 we present an adaptive tester for Clique-Collection that has almost-linear query complexity. This result stands in contrast to the (tight) lower-bound on the query complexity of non-adaptive testers for Clique-Collection, presented in Section 4. Larger gaps between the query complexity of adaptive versus non-adaptive testers are presented in Section 5. On the other hand, in Section 6, we present non-adaptive testers of query complexity that is almost-linear in the reciprocal of the proximity parameter.

2 Preliminaries

In this section we review the definition of property testing, when specialized to graph properties in the adjacency matrix model. We also define several natural graph properties, which will serve as the pivot of our study.

2.1 Basic notions

For an integer n, we let $[n] = \{1, ..., n\}$. A generic N-vertex graph is denoted by $G = ([N], E)$, where $E \subseteq \{\{u, v\} : u, v \in [N]\}$ is a set of (unordered) pairs of vertices. Any set of (such) graphs that is closed under isomorphism is called a graph property. By oracle access to such a graph $G = ([N], E)$ we mean oracle access to the Boolean function that answers the query $\{u, v\}$ (or rather $(u, v) \in [N] \times [N]$) with the bit 1 if and only if $\{u, v\} \in E$.

Definition 2.1 (property testing for graphs in the adjacency matrix model): *A* tester *for a graph property Π is a probabilistic oracle machine that, on input parameters N and ϵ and access to an N-vertex graph $G = ([N], E)$, output a binary verdict that satisfies the following two conditions.*

1. *If $G \in \Pi$ then the tester accepts with probability at least* $2/3$.
2. *If G is ϵ-far from Π then the tester accepts with probability at most* $1/3$, *where G is* ϵ-far *from Π if for every N-vertex graph $G' = ([N], E') \in \Pi$ it holds that the symmetric difference between E and E' has cardinality at least ϵN^2.*[6]

If the tester accepts every graph in Π with probability 1, then we say that it has one-sided error. *A tester is called* non-adaptive *if it determines all its queries based solely on its internal coin tosses* (and the parameters N and ϵ)*; otherwise it is called* adaptive.

[6]Indeed, it is more natural to require that this symmetric difference should have cardinality at least $\epsilon \cdot \binom{N}{2}$. The current convention is adopted for sake of convenience.

The query complexity of a tester is the number of queries it makes to any N-vertex graph oracle, as a function of the parameters N and ϵ. We say that a tester is efficient if it runs in time that is polynomial in its query complexity, where basic operations on elements of $[N]$ are counted at unit cost. We note that all testers presented in this paper are efficient, whereas the lower-bounds hold also for non-efficient testers.

We shall focus on properties that can be tested within query complexity that only depends on the proximity parameter, ϵ. Thus, the query-complexity upper-bounds that we state hold for any values of ϵ and N, but will be meaningful only for $\epsilon > 1/N^2$ or so. In contrast, the lower-bounds (e.g., of $\Omega(1/\epsilon)$) cannot possibly hold for $\epsilon < 1/N^2$, but they will indeed hold for any $\epsilon > N^{-\Omega(1)}$. Alternatively, one may consider the query-complexity as a function of ϵ, where for each fixed value of $\epsilon > 0$ the value of N tends to infinity.

Notation and a convention. For a fixed graph $G = ([N], E)$, we denote by $\Gamma(v) = \{u\colon\{u, v\} \in E\}$ the set of neighbors of vertex v. At times, we look at E as a subset of $V \times V$; that is, we often identify E with $\{(u, v)\colon\{u, v\}\in E\}$. If a graph $G = ([N], E)$ is not ϵ-far from a property Π then we say that G is ϵ-close to Π; this means that less than ϵN^2 edges should be added and/or removed from G such to yield a graph in Π.

2.2 The graph properties to be studied

The set of graphs that consists of a collection of isolated cliques is called clique collection and is denoted $\mathcal{CC}$; that is, a graph $G = ([N], E)$ is in $\mathcal{CC}$ if and only if the vertex set $[N]$ can be partitioned into $(C_1, ..., C_t)$ such that the subgraph induced by each C_i is a clique and there are no edges with endpoints in different C_i's (i.e., for every $u < v \in [N]$ it holds that $\{u, v\} \in E$ if and only if there exists an i such that $u, v \in C_i$). If $t \leq c$ then we say that G is in $\mathcal{CC}^{\leq c}$; that is, $\mathcal{CC}^{\leq c}$ is the subset of $\mathcal{CC}$ that contains graphs that are each a collection of up-to c isolated cliques.

A bi-clique is a complete bipartite graph (i.e., a graph $G = (V, E)$ such that V is partitioned into $(S, V \setminus S)$ such that $\{u, v\} \in E$ if and only if $u \in S$ and $v \in V \setminus S$). Note that a graph is a bi-clique if and only if its complement is in $\mathcal{CC}^{\leq 2}$. The set of graphs that consists of a collection of isolated bi-cliques is called bi-clique collection and denoted $\mathcal{BCC}$; that is, a graph $G = ([N], E)$ is in $\mathcal{BCC}$ if and only if the vertex set $[N]$ can be partitioned into $(V_1, ..., V_t)$ such that the subgraph induced by each V_i is a bi-clique and there are no edges with endpoints in different V_i's (i.e., each V_i is partitioned into $(S_i, V_i \setminus S_i)$ such that for every $u < v \in [N]$ it holds that $\{u, v\} \in E$ if and only if there exists an i such that $(u, v) \in S_i \times (V \setminus S))$.

Generalizations of $\mathcal{BCC}$ are obtained by considering collections of "super-paths" and "super-cycles" respectively. A super-path (of length t) is a sequence of disjoint sets of vertices, $S_1, ..., S_t$, such that vertices $u, v \in \bigcup_{i\in[t]} S_i$ are connected by an edge if and only if for some $i \in [t-1]$ it holds that $u \in S_i$ and $v \in S_{i+1}$. Note that a bi-clique can be viewed as a super-path of length two. We denote the set of graphs that consists of a collection of isolated super-paths of length t by $\mathcal{S}\mathsf{P}_t\mathcal{C}$ (e.g., $\mathcal{S}\mathsf{P}_2\mathcal{C} = \mathcal{BCC}$). Similarly, a super-cycle (of length t) is a sequence of disjoint sets of vertices, $S_1, ..., S_t$, such that vertices $u, v \in \bigcup_{i\in[t]} S_i$ are connected by an edge if and only if for some $i \in [t]$ it holds that $u \in S_i$ and $v \in S_{(i \bmod t)+1}$. Note that a bi-clique that has at least two vertices on each side can be viewed as a super-cycle of length four (by partitioning each of its sides into two parts). We denote the set of graphs that consists of a collection of isolated super-cycles of length t by $\mathcal{S}\mathsf{C}_t\mathcal{C}$ (e.g., $\mathcal{S}\mathsf{C}_4\mathcal{C} \subset \mathcal{BCC}$, where the strict containment is due to the pathological case of bi-cliques having at most one node on one side).

2.3 Annoying technicalities

We allowed ourselves various immaterial inaccuracies. For example, various quantities (e.g., $\log_2(1/\epsilon)$) are treated as if they are integers, whereas one should actually use some rounding and compensate for the rounding error. At times, we ignore events that occur with probability that is inversely proportional to the number of vertices; for example, when we select a random sample of $s = O(1)$ (or $s = \widetilde{O}(1/\epsilon)$) vertices, we often analyze it as if sampling was done with repetitions. In some places, we do not specify the "high" (constant) probability with which some events occur; but such missing details are easy to fill-up. In other places, we specify high constants that are not the best ones possible.

3 The Adaptive Query Complexity of Clique-Collection

In this section we study the (adaptive) query complexity of clique collection, presenting an almost optimal (adaptive) tester for this property. Loosely speaking, the tester starts by finding a few random neighbors of a few randomly selected start vertices, and then examines the existence of edges among the neighbors of each start vertex as well as among these neighbors and the non-neighbors of each start vertex.

We highlight the fact that adaptivity is used in order to perform queries that refer only to pairs of neighbors of the same start vertex. To demonstrate the importance of this fact, consider the case that the N-vertex graph is partitioned into $O(1/\epsilon)$ connected components each having $O(\epsilon N)$ vertices. Suppose that we wish to tell whether the connected component that contains the vertex v is indeed a clique. Using adaptive queries we may first find two neighbors of v, by selecting $t \stackrel{\rm def}{=} O(1/\epsilon)$ random vertices and checking whether each such vertex is adjacent to v, and then check whether these *two* neighbors are adjacent. In contrast, intuitively, a non-adaptive procedure cannot avoid making all $\binom{t}{2}$ possible queries.

The foregoing adaptive procedure is tailored to the case that the N-vertex graph is partitioned into $O(1/\epsilon)$ ("strongly connected") components, each having $O(\epsilon N)$ vertices. In such a case, it suffices to check that a constant fraction of these components are in fact cliques (or rather close to being so) and that there are no edges (or rather relatively few edges) from these cliques to the rest of the graph. However, if the components (and potential cliques) are larger, then we should check more of them, but (fortunately) due to their larger size finding neighbors requires less queries, and the total number of queries remains invariant. These considerations lead us to the following algorithm.

Algorithm 3.1 (adaptive tester for $\mathcal{CC}$): *On input N and ϵ and oracle access to a graph $G = ([N], E)$, the tester sets $t_1 = \Theta(1)$ and $t_2 = \Theta(\log^3(1/\epsilon))$, and proceeds in $\ell \stackrel{\rm def}{=} \log_2(1/\epsilon) + 2$ iterations as follows: For $i = 1, ..., \ell$, the tester selects uniformly $t_1 \cdot 2^i$ start vertices and for each selected vertex $v \in [N]$ performs the following sub-test, denoted* sub-test$_i(v)$*:*

1. *The sub-test selects at random a sample, S, of $t_2/(2^i\epsilon)$ vertices.*
2. *The sub-test determines $\Gamma_S(v) = S \cap \Gamma(v)$, by making the queries (v, w) for each $w \in S$.*
3. *If $|\Gamma_S(v)| \leq \sqrt{t_2/2^i\epsilon}$ then the sub-test checks that for every $u, w \in \Gamma_S(v)$ it holds that $(u, w) \in E$. Otherwise* (i.e., $|\Gamma_S(v)| > \sqrt{t_2/2^i\epsilon}$)*, it selects a sample of $t_2/(2^i\epsilon)$ pairs in $\Gamma_S(v) \times \Gamma_S(v)$ and checks that each selected pair is in E.*

4. *The sub-test selects a sample of $t_2/(2^i\epsilon)$ pairs in $\Gamma_S(v) \times (S \setminus \Gamma_S(v))$ and checks that each selected pair is* not *in E.*

The sub-test (i.e., sub-test$_i(v)$) *accepts if and only if all checks were positive* (i.e., no edges were missed in Step 3 and no edges were detected in Step 4). *The tester itself accepts if and only if all $\sum_{i=1}^{\ell} t_1 \cdot 2^i$ invocations of the sub-test accepted.*

The query complexity of this algorithm is $\sum_{i=1}^{\ell} t_1 2^i \cdot O(t_2/2^i\epsilon) = O(\ell \cdot t_1 t_2/\epsilon) = \widetilde{O}(1/\epsilon)$, and evidently it is efficient. Clearly, this algorithm accepts (with probability 1) any graph that is in $\mathcal{CC}$. It remains to analyze its behavior on graphs that are ϵ-far from $\mathcal{CC}$.

Lemma 3.2 *If $G = ([N], E)$ is ϵ-far from $\mathcal{CC}$, then on input N, ϵ and oracle access to G, Algorithm 3.1 rejects with probability at least $2/3$.*

Part 1 of Theorem 1.1 follows.

Proof: We shall prove the contrapositive statement; that is, that if Algorithm 3.1 accepts with probability at least 1/3 then the graph is ϵ-close to $\mathcal{CC}$. The proof evolves around the following notion of i-good start vertices (for $i \in [\ell]$). We shall first show that if Algorithm 3.1 accepts with probability at least 1/3 then the number of "important" vertices that are not i-good is relatively small, and next show how to use the i-good vertices in order to construct a partition of the vertices that demonstrates that the graph is ϵ-close to $\mathcal{CC}$. The following definition refers to a parameter γ_2, which will be set to $\Theta(1/t_2)$.

Definition 3.2.1 *A vertex v is i-*good *if the following two conditions hold.*

1. *The subgraph induced by $\Gamma(v)$ misses at most $\gamma_2 \cdot 2^i\epsilon \cdot |\Gamma(v)| \cdot N$ edges.*
2. *For every positive integer $j \le j_0 \stackrel{\text{def}}{=} \log_2(|\Gamma(v)|/(\gamma_2 \cdot 2^i\epsilon N))$, the number of vertices in $\Gamma(v)$ that have at least $\gamma_2 \cdot 2^{i+j}\epsilon \cdot N$ edges going out of $\Gamma(v)$ is at most $2^{-j} \cdot |\Gamma(v)|$.*

Note that Condition 1 holds vacuously whenever $|\Gamma(v)| < \gamma_2 \cdot 2^i\epsilon \cdot N$. However, when $|\Gamma(v)| \gg \gamma_2 \cdot 2^i\epsilon \cdot N$, Condition 1 implies that at least 99% of the vertices in $\Gamma(v)$ have at least $0.99 \cdot |\Gamma(v)|$ neighbors in $\Gamma(v)$. Condition 2 implies that, when ignoring at most $2^{-j_0} \cdot |\Gamma(v)| < \gamma_2 \cdot 2^i\epsilon \cdot N$ vertices (in $\Gamma(v)$), the number of edges going out of $\Gamma(v)$ is at most $\sum_{j=1}^{j_0} 2^{-(j-1)}|\Gamma(v)| \cdot \gamma_2 2^{i+j}\epsilon N$, which is less than $4\ell \cdot \gamma_2 2^i\epsilon \cdot |\Gamma(v)| \cdot N$, since $j_0 \le \log_2(1/\gamma_2 2^i\epsilon) \le \log_2(1/\gamma_2\epsilon) < 2\log_2(1/\epsilon)$.

Claim 3.2.2 *If v has degree at least $\gamma_2 \cdot 2^i\epsilon \cdot N$ and is not i-good, then the probability that* sub-test$_i(v)$ *accepts is less than 5%.*

Proof: Intuitively, the lower-bound on $|\Gamma(v)|$ implies that the violation of any of the two conditions of Definition 3.2.1 is detected with high probability by sub-test$_i(v)$. For example, if 1% of the vertices in $\Gamma(v)$ have less than $0.99 \cdot |\Gamma(v)|$ neighbors in $\Gamma(v)$, then the residual sample $\Gamma_S(v)$ (created by sub-test$_i(v)$) is likely to contain a constant fraction of vertices that miss a constant fraction of neighbors in $\Gamma_S(v)$. The actual proof, which refers to the two conditions of i-goodness, follows.

Assume that Condition 1 of i-goodness does not hold for v, and let $\rho \stackrel{\text{def}}{=} \frac{\gamma_2 \cdot 2^i\epsilon \cdot |\Gamma(v)| \cdot N}{|\Gamma(v)|^2} = \frac{\gamma_2 \cdot 2^i\epsilon \cdot N}{|\Gamma(v)|}$ denote (the lower bound on) the fraction of missing edges in $\Gamma(v)$. (Note that this event may happen only if $|\Gamma(v)| \ge \gamma_2 \cdot 2^i\epsilon \cdot N$.) Then, with probability at least 0.9, it holds that $|\Gamma_S(v)| > m/2$, where $m \stackrel{\text{def}}{=} \frac{t_2}{\epsilon 2^i} \cdot \frac{|\Gamma(v)|}{N} \geq t_2 \cdot \gamma_2 \gg 1$. Also note that the members of $\Gamma_S(v)$ are distributed uniformly in $\Gamma(v)$. Now, consider $n = m/2$ uniformly distributed vertices in $\Gamma(v)$, and let $\zeta_{i,j} = 1$ if there is no edge between the i^{th} and j^{th} vertices in the sample. Then, $\text{Exp}(\zeta_{i,j}) \geq \rho$. Applying Chebyshev's Inequality[7] it follows that, with probability at least 0.9, the fraction of edges that are missing in the subgraph induced by the said sample is at least $\rho/2$. It follows that Step 3 of sub-test$_i(v)$ rejects with probability at least 0.9^2 (regardless if it examines all pairs in $\Gamma_S(v) \times \Gamma_S(v)$ or just examines a random sample of $\frac{t_2}{2^i \epsilon} \geq \frac{t_2 \gamma_2}{\rho}$ pairs).

Assume that Condition 2 of i-goodness does not hold for v; that is, there exists a $j \leq j_0$ such that more than $2^{-j} \cdot |\Gamma(v)|$ vertices in $\Gamma(v)$ have each at least $\gamma_2 \cdot 2^{i+j}\epsilon \cdot N$ edges going out of $\Gamma(v)$. Using the same setting of m and n as in the previous paragraph (as well as the hypothesis $|\Gamma(v)| \geq \gamma_2 \cdot 2^i \epsilon \cdot N$), we note (again) that with high probability $|\Gamma_S(v)| > n$, and that $\Gamma_S(v)$ is expected to contain $n \cdot 2^{-j} = t_2 \gamma_2 \cdot 2^{j_0 - j} \geq t_2 \gamma_2$ vertices of "high out-degree" (and it will contain approximately such a number, with high probability). It follows that the number of pairs in $\Gamma_S(v) \times ([N] \setminus \Gamma(v))$ that are edges is at least $n 2^{-j} \cdot \gamma_2 \cdot 2^{i+j} \epsilon N/2$, which means an edge density of at least $\rho' \stackrel{\text{def}}{=} \gamma_2 \cdot 2^i \epsilon/2$. Since $|S| = \frac{t_2}{2^i \epsilon} \gg 1/\rho'$, with high probability, approximately the same edge density is maintained also in $\Gamma_S(v) \times (S \setminus \Gamma_S(v))$. Thus, a sample of $\frac{t_2}{2^i \epsilon}$ random pairs in $\Gamma_S(v) \times (S \setminus \Gamma_S(v))$ will hit an edge with high probability and cause Step 4 (of sub-test$_i(v)$) to reject. The claim follows. $\square$

Claim 3.2.3 *If Algorithm 3.1 accepts with probability at least* $1/3$ *then for every* $i \in [\ell]$ *the number of vertices of degree at least* $\gamma_2 \cdot 2^i \epsilon \cdot N$ *that are not* i*-good is at most* $\gamma_1 \cdot 2^{-i} \cdot N$, *where* $\gamma_1 \stackrel{\text{def}}{=} \Theta(1/t_1)$.

Claim 3.2.3 follows by combining Claim 3.2.2 with the fact that Algorithm 3.1 invokes sub-test$_i$ on $t_1 \cdot 2^i$ random vertices (and using $(1 - \gamma_1 \cdot 2^{-i})^{t_1 \cdot 2^i} + 0.05 < 1/3$). Next, using the conclusion of Claim 3.2.3, we turn to construct a partition $(C_1, ..., C_t)$ of $[N]$ such that the graph G misses at most $\epsilon \cdot N^2/2$ edges within the C_i's and has at most $\epsilon \cdot N^2/2$ edges between the C_i's. The partition is constructed in iterations. *We start with a motivating discussion.*

Note that any i-good vertex, v, yields a set of vertices (i.e., $\Gamma(v)$) that is "close" to being a clique, where "closeness" has a stricter meaning when i is smaller. Specifically, by Condition 1, this set misses at most $\gamma_2 \cdot 2^i \epsilon \cdot |\Gamma(v)| \cdot N$ edges. But we should also care about how this set "interacts" with the rest of the graph, which is where Condition 2 comes into play. Letting C_v contain only the vertices in $\Gamma(v)$ that have less than $|\Gamma(v)|$ neighbors outside of $\Gamma(v)$, we upper-bound the number of edges going out of C_v as follows: We first note that these edges are either edges between C_v and $\Gamma(v) \setminus C_v$ or edges between C_v and $[N] \setminus \Gamma(v)$. The number of edges of the first type is upper-bounded by $|C_v| \cdot |\Gamma(v) \setminus C_v|$, which (by using Condition 2 and $j_0 = \log_2(|\Gamma(v)|/(\gamma_2 \cdot 2^i \epsilon N)))$ is upper-bounded by $|C_v| \cdot 2^{-j_0} |\Gamma(v)| = |C_v| \cdot \gamma_2 2^i \epsilon N \leq \gamma_2 2^i \epsilon \cdot |\Gamma(v)| \cdot N$. The number of edges of the second type is upper-bounded by

$$\sum_{j=1}^{j_0} 2^{-(j-1)} |\Gamma(v)| \cdot \gamma_2 \cdot 2^{i+j} \epsilon \cdot N \;=\; 2 j_0 \cdot \gamma_2 2^i \epsilon \cdot |\Gamma(v)| \cdot N, \qquad (1)$$

[7] Here we have $\binom{n}{2}$ random variables, which are partially pairwise independent (i.e., $\zeta_{i,j}$ is independent of $\zeta_{i',j'}$ if $|\{i,j,i',j'\}| = 4$). Furthermore, these random variables assume values in $\{0,1\}$ (and so $\zeta_{i,j}^2 = \zeta_{i,j}$) and it holds that $n \cdot \rho = t_2 \gamma_2/2 \gg 1$ (rather than merely $n^2 \gg 1/\rho$). Assume, for simplicity that $\text{Exp}(\zeta_{i,j}) = \rho$. It follows that $\text{Exp}(\sum_{i<j} \zeta_{i,j}) = \binom{n}{2} \cdot \rho > n^2 \rho/3$ and $\text{Var}(\sum_{i<j} \zeta_{i,j}) < 4 \cdot \text{Exp}(\sum_{i<j,k} \zeta_{i,j} \zeta_{i,k}) = 4n \cdot \text{Exp}(\sum_{i<j} \zeta_{i,j}) < 2n^3 \rho$. Thus, $\frac{\text{Var}}{\text{Exp}^2} < \frac{18}{n\rho} = \frac{36}{t_2 \gamma_2}$, which can be made an arbitrary small constant (by an adequate choice of $t_2 = \Theta(1/\gamma_2)$).

by assigning each vertex $u \in C_v$ the smallest $j \in [j_0]$ such that $|\Gamma(u) \setminus \Gamma(v)| < \gamma_2 \cdot 2^{i+j}\epsilon \cdot N$, and using $\gamma_2 2^{i+j_0}\epsilon \cdot N = |\Gamma(v)|$. Thus, the total number of these edges is upper-bounded by $(2j_0+1)\cdot\gamma_2 2^i\epsilon\cdot|\Gamma(v)|\cdot N$, which is upper-bounded by $3\ell\cdot\gamma_2 2^i\epsilon\cdot|\Gamma(v)|\cdot N$ (since $j_0 \leq \log_2(1/(\gamma_2\cdot 2^i\epsilon)) \leq \log_2(1/\gamma_2\epsilon) = (1+o(1))\cdot\ell$).

The foregoing paragraph identifies a single (good) clique, while we wish to identify all cliques. Starting with $i = 1$, the basic idea is identifying new cliques by using i-good vertices that are not covered by previously identified cliques. If we are lucky and the entire graph is covered this way then we halt. But it may indeed be the case that some vertices are left uncovered and that they are not i-good. At this point we invoke Claim 3.2.3 and conclude that these vertices either have low degree (i.e., have degree at most $\gamma_2 \cdot 2^i\epsilon \cdot N$) or are relatively few in number (i.e., their number is at most $\gamma_1 \cdot 2^{-i} \cdot N$). Ignoring (for a moment) the vertices of low degree, we deal with the remaining vertices by invoking the same reasoning with respect to an incremented value of i (i.e., $i \leftarrow i+1$). The key observation is that the number of violations, caused by cliques identified in each iteration i, is upper-bounded by the product of the number of vertices covered in that iteration (which is linearly related to 2^{-i}) and the "density" of violations caused by each identified clique (which is linearly related to $2^i\epsilon$). Thus, intuitively, each iteration contributes $O(\ell\gamma_2\epsilon \cdot N^2)$ violations, and after the last iteration (i.e., $i = \ell$) we are left with at most $\gamma_1 \cdot 2^{-i} \cdot N < \gamma_1\epsilon N$ vertices, which we can afford to identify as a single clique (or alternatively as isolated vertices).

Two problems, which were ignored by the foregoing description, arise from the fact that vertices that are identified as belonging to the clique C_v (of some i-good vertex v) may belong either to previously identified cliques or to the set of vertices cast aside as having low degree. Our solution is using only i-good vertices for which the majority of neighbors do not belong to these two categories (i.e., vertices v such that most of $\Gamma(v)$ belongs neither to previously identified cliques nor have low degree). This leads to the following description.

The partition reconstruction procedure. The iterative procedure is initiated with $C = L_0 = \emptyset$, $R_0 = [N]$ and $i = 1$, where C denotes the set of vertices "covered" (by cliques) so far, R_{i-1} denotes the set of "remaining" vertices after iteration $i-1$ and L_{i-1} denotes the set of vertices cast aside (as having "low degree") in iteration $i-1$. The procedure refers to a parameter $\beta = \Theta(1/\ell) \gg \gamma_2$, which determines the "low degree" threshold (for each iteration). The i^{th} iteration proceeds as follows, where $i = 1, ..., \ell$ and F_i is initialized to $\emptyset$.

1. Pick an arbitrary vertex $v \in R_{i-1} \setminus C$ that satisfies the following three conditions

   (a) v is i-good.

   (b) v has sufficiently high degree; that is, $|\Gamma(v)| \geq \beta \cdot 2^i\epsilon \cdot N$.

   (c) v has relatively few neighbors in C; that is, $|\Gamma(v) \cap C| \leq |\Gamma(v)|/4$.

   If no such vertex exists, define $L_i = \{v \in R_{i-1} \setminus C : |\Gamma(v)| < \beta\cdot 2^i\epsilon\cdot N\}$ and $R_i = R_{i-1} \setminus (L_i \cup C)$. If $i < \ell$ then proceed to the next iteration, and otherwise terminate.

2. For vertex v as selected in Step 1, let $C_v = \{u \in \Gamma(v) : |\Gamma(u) \setminus \Gamma(v)| < |\Gamma(v)|\}$. Form a new clique with the vertex set $C'_v \leftarrow C_v \setminus C$, and update $F_i \leftarrow F_i \cup \{v\}$ and $C \leftarrow C \cup C'_v$.

Note that by Condition 1c, for every $v \in F_i$, it holds that $|C'_v| \geq |C_v| - (|\Gamma(v)|/4)$, whereas by i-goodness[8] (and $j_0 = \log_2(|\Gamma(v)|/(\gamma_2 \cdot 2^i\epsilon N)) \geq \log_2(\beta/\gamma_2) = \omega(1)$) we have $|C_v| > (1-o(1))\cdot|\Gamma(v)|$.

[8] Every $v \in F_i$ is i-good and thus satisfies $|C_v| > (1 - 2^{-j_0}) \cdot |\Gamma(v)|$.

Thus, quality guarantees that are quantified in terms of $|\Gamma(v)|$ translate well to similar guarantees in terms of $|C'_v|$. This fact, combined with the fact that C_v cannot contain many low degree vertices (i.e., vertices cast aside (in prior iterations) as having low degree), plays an important role in the following analysis.

Claim 3.2.4 *Referring to the partition reconstruction procedure, for every $i \in [\ell]$, the following holds.*

1. *The number of missing edges inside the cliques formed in iteration i is at most $8\gamma_2\epsilon \cdot N^2$; that is,*

$$\left|\bigcup_{v\in F_i} \{(u,w) \in C'_v \times C'_v : (u,w) \notin E\}\right| \leq 8\gamma_2\epsilon \cdot N^2.$$

2. *The number of* ("superfluous") *edges between cliques formed in iteration i and either R_i or other cliques formed in the same iteration is $24\ell \cdot \gamma_2\epsilon \cdot N^2$; actually,*

$$\left|\bigcup_{v\in F_i} \{(u,w) \in C'_v \times (R_{i-1} \setminus C'_v) : (u,w) \in E\}\right| \leq 24\ell \cdot \gamma_2\epsilon \cdot N^2.$$

3. $|R_i| \leq 2^{-i} \cdot N$ *and* $|L_i| \leq 2^{-(i-1)} \cdot N$.

Thus, the total number of violations caused by the cliques that are formed by the foregoing procedure is upper-bounded by $(24+o(1))\ell^2 \cdot \gamma_2\epsilon \cdot N^2 = o(\epsilon N^2)$. (We mention that the setting $\gamma_2 = o(\ell^2)$ is used for establishing Item 3.)

Proof: We prove all items simultaneously, by induction from $i = 0$ to $i = \ell$. Needless to say, all items hold vacuously for $i = 0$, and thus we focus on the induction step.

Starting with Item 1, we note that every $v \in F_i$ is i-good and thus the number of edges missing in $C'_v \times C'_v \subseteq \Gamma(v) \times \Gamma(v)$ is at most $\gamma_2 2^i\epsilon \cdot |\Gamma(v)| \cdot N < 2\gamma_2 2^i\epsilon \cdot |C'_v| \cdot N$, where the inequality follows from $|C'_v| > |\Gamma(v)|/2$ (which follows by combining $|C'_v| \geq |C_v| - (\Gamma(v)|/4)$ and $|C_v| \geq (1 - 2^{-j_0}) \cdot |\Gamma(v)|$, where $j_0 = \log_2(|\Gamma(v)|/(\gamma_2 \cdot 2^i\epsilon N)) > 2$). Recall that the i-goodness of v (combined with $|\Gamma(v)| \geq \beta \cdot 2^i\epsilon \cdot N$) implies that $\Gamma(v)$ contains at least $0.99 \cdot |\Gamma(v)|$ vertices of degree exceeding $0.99 \cdot |\Gamma(v)|$. This implies that $|\Gamma(v) \cap (\bigcup_{j\in[i-1]} L_j)| < |C_v|/4$, because $|\Gamma(v)| \geq \beta 2^i\epsilon \cdot N$ whereas every vertex in $\bigcup_{j\in[i-1]} L_j$ has degree at most $\beta 2^{i-1}\epsilon \cdot N$. Observing that $C'_v = (C'_v \cap R_{i-1}) \cup (C'_v \cap \bigcup_{j\in[i-1]} L_j)$, it follows that $|\bigcup_{v\in F_i} C'_v \cap R_{i-1}| > |\bigcup_{v\in F_i} C'_v|/2$, and thus $\sum_{v\in F_i} |C'_v| \leq 2|R_{i-1}|$. Combining all these bounds, we obtain

$$\begin{aligned}
\left|\bigcup_{v\in F_i} \{(u,w) \in C'_v \times C'_v : (u,w) \notin E\}\right| &= \sum_{v\in F_i} |\{(u,w) \in C'_v \times C'_v : (u,w) \notin E\}| \\
&\leq 2\gamma_2 2^i\epsilon \cdot \sum_{v\in F_i} |C'_v| \cdot N \\
&\leq 2\gamma_2 2^i\epsilon \cdot 2|R_{i-1}| \cdot N.
\end{aligned}$$

Using the induction hypothesis regarding R_{i-1} (i.e., $|R_{i-1}| < 2^{-(i-1)} \cdot N$), Item 1 follows.

Item 2 is proved in a similar fashion. Here we use the fact[9] that i-goodness of v (which follows from $v \in F_i$) implies that the number of edges in $C'_v \times (R_{i-1} \setminus C'_v) \subseteq C_v \times ([N] \setminus C_v)$ is at most $3\ell \cdot \gamma_2 2^i \epsilon \cdot |\Gamma(v)| \cdot N$, which is upper-bounded by $6\ell \cdot \gamma_2 2^i \epsilon \cdot |C'_v| \cdot N$. Using again $\sum_{v \in F_i} |C'_v| < 2|R_{i-1}|$ and $|R_{i-1}| < 2^{-(i-1)} \cdot N$, we establish Item 2.

Turning to Item 3, we first note that $L_i \subseteq R_{i-1}$ and thus $|L_i| \leq |R_{i-1}| \leq 2^{-(i-1)} \cdot N$. As for R_i, it may contain only vertices that are neither in L_i nor in $\bigcup_{v \in F_i} C'_v$. It follows that for every $v \in R_i$ either v is not i-good (although it has degree at least $\beta \cdot 2^i \epsilon \cdot N$) or it has at least $|\Gamma(v)|/4$ neighbors in previously identified cliques (which implies $|\Gamma(v) \cap (\bigcup_{w \in \bigcup_{j \in [i]} F_j} C'_w)| \geq |\Gamma(v)|/4$). By Claim 3.2.3, the number of vertices of the first type is at most $\gamma_1 2^{-i} \cdot N$. As for vertices of the second type, each such vertex v (in R_i) requires at least $|\Gamma(v)|/4 \geq \beta \cdot 2^i \epsilon \cdot N/4$ edges from $C' \stackrel{\text{def}}{=} \bigcup_{w \in \bigcup_{j \in [i]} F_j} C'_w$ to it (because C' is the set of vertices covered by previously identified cliques at the time iteration i is completed). By Item 2, the total number of edges going out from C' to R_i is at most $i \cdot 24\ell \cdot \gamma_2 \epsilon \cdot N^2 \leq 24\ell^2 \cdot \gamma_2 \epsilon \cdot N^2$. On the other hand, as noted above, each vertex of the second type has at least $\beta \cdot 2^i \epsilon \cdot N/4$ edges incident to vertices in C'. Hence, the number of vertices of the second type is upper-bounded by

$$\frac{24\ell^2 \cdot \gamma_2 \epsilon \cdot N^2}{\beta \cdot 2^i \epsilon \cdot N} = \frac{24\ell^2 \cdot \gamma_2}{\beta} \cdot 2^{-i} N, \tag{2}$$

Thus, $|R_i| \leq (\gamma_1 + 24\ell^2 \gamma_2 \beta^{-1}) \cdot 2^{-i} \cdot N$. By the foregoing setting of γ_1, γ_2 and β (e.g., $\gamma_1 = 1/2$ and $\gamma_2 = \beta/(48\ell^2)$), it follows that $|R_i| \leq 2^{-i} \cdot N$. □

Completing the reconstruction and its analysis. The foregoing construction leaves "unassigned" the vertices in R_ℓ as well as some of the vertices in $L_1, ..., L_\ell$. (Note that some vertices in $\bigcup_{i=1}^{\ell-1} L_i$ may be placed in cliques constructed in later iterations, but there is no guarantee that this actually happens.) We now assign each of these remaining vertices to a singleton clique (i.e., an isolated vertex). The number of violation caused by this assignment equals the number of edges with both endpoints in $R' \stackrel{\text{def}}{=} R_\ell \cup \bigcup_{i=1}^{\ell} L_i$, because edges with a single endpoint in R' were already accounted for in Item 2 of Claim 3.2.4. Nevertheless, we upper-bound the number of violations by the total number of edges adjacent at R', which in turn is upper-bounded by

$$\begin{aligned}
\sum_{v \in R_\ell \cup \bigcup_{i \in [\ell]} L_i} |\Gamma(v)| &\leq |R_\ell| \cdot N + \sum_{i=1}^{\ell} \sum_{v \in L_i} |\Gamma(v)| \\
&\leq \frac{\epsilon N}{4} \cdot N + \sum_{i=1}^{\ell} 2^{-(i-1)} N \cdot \beta 2^i \epsilon N \\
&= \frac{\epsilon}{4} \cdot N^2 + 2\ell \cdot \beta \cdot \epsilon N^2.
\end{aligned}$$

By the foregoing setting of β (i.e., $\beta \leq 1/8\ell$), it follows that the number of these edges is smaller than $\epsilon N^2/2$. Combining this with the bounds on the number of violating edges (or non-edges) as provided by Claim 3.2.4, the lemma follows. ■

[9] This fact was established in the motivating discussion that precedes the description of the procedure (see Eq. (1) and its vicinity). Specifically, recall that the number of edges in $C_v \times ([N] \setminus C_v)$ is upper-bounded by the sum of $|C_v \times (\Gamma(v) \setminus C_v)|$ and the number of edges in $C_v \times ([N] \setminus \Gamma(v))$. Using Condition 2 of i-goodness, we upper-bound both $|\Gamma(v) \setminus C_v|$ and the number of edges of the second type, and the fact follows.

4 The Non-Adaptive Query Complexity of Clique-Collection

In this section we study the non-adaptive query complexity of clique collection. We first establish the lower-bound claimed in Part 2 of Theorem 1.1, and next show that this lower-bound is essentially tight.

4.1 The Lower Bound

In this section we establish Part 2 of Theorem 1.1. Specifically, for every value of $\epsilon > 0$, we consider two different sets of graphs, one consisting of graphs in $\mathcal{CC}$ and the other consisting of graphs that are ϵ-far from $\mathcal{CC}$, and show that a non-adaptive algorithm of query complexity $o(\epsilon^{-4/3})$ cannot distinguish between graphs selected at random in these sets.

The first set, denoted $\mathcal{CC}_\epsilon$, contains all N-vertex graphs such that each graph consists of $(2\epsilon)^{-1}$ cliques, and each clique has size $2\epsilon \cdot N$. It will be instructive to partition these $(2\epsilon)^{-1}$ cliques into $(4\epsilon)^{-1}$ pairs (each consisting of two cliques). The second set, denoted $\mathcal{BCC}_\epsilon$, contains all N-vertex graphs such that each graph consists of $(4\epsilon)^{-1}$ bi-cliques, and each bi-clique has $2\epsilon \cdot N$ vertices on each side. Indeed, $\mathcal{CC}_\epsilon \subseteq \mathcal{CC}$, whereas each graph in $\mathcal{BCC}_\epsilon$ is ϵ-far from $\mathcal{CC}$ (because each of the bi-cliques must be turned into a collection of cliques).

In order to motivate the claim that a non-adaptive algorithm of query complexity $o(\epsilon^{-4/3})$ cannot distinguish between graphs selected at random in these sets, consider the (seemingly best such) algorithm that selects $o(\epsilon^{-2/3})$ vertices and inspects the induced subgraph. Consider the partition of a graph in $\mathcal{CC}_\epsilon$ into $(4\epsilon)^{-1}$ pairs of cliques, and correspondingly the partition of a graph in $\mathcal{BCC}_\epsilon$ into $(4\epsilon)^{-1}$ bi-cliques. Then, the probability that a sample of $o(\epsilon^{-2/3})$ vertices contains at least three vertices that reside in the same part (of $4\epsilon \cdot N$ vertices) is $o(\epsilon^{-2/3})^3 \cdot (4\epsilon)^2 = o(1)$. On the other hand, if this event does not occur, then the answers obtained from both graphs are indistinguishable (because in each case a random pair of vertices residing in the same part is connected by an edge with probability $1/2$). As will be shown below, this intuition extends to an arbitrary non-adaptive algorithm.

Specifically, by an averaging argument, it suffices to consider deterministic algorithms, which are fully specified by the sequence of queries that they make and their decision on each corresponding sequence of answers. Recall that these (fixed) queries are elements of $[N] \times [N]$. We shall show that, for every sequence of $o(\epsilon^{-4/3})$ queries, the answers provided by a randomly selected element of $\mathcal{CC}_\epsilon$ are statistically close to the answers provided by a randomly selected element of $\mathcal{BCC}_\epsilon$. We shall use the following notation: For an N-vertex graph G and a query (u, v), we denote the corresponding answer by $\mathrm{ans}_G(u, v)$; that is, $\mathrm{ans}_G(u, v) = 1$ if $\{u, v\}$ is an edge in G and $\mathrm{ans}_G(u, v) = 0$ otherwise.

Lemma 4.1 *Let G_1 and G_2 be random N-vertex graphs uniformly distributed in $\mathcal{CC}_\epsilon$ and $\mathcal{BCC}_\epsilon$, respectively. Then, for every sequence $(v_1, v_2), ..., (v_{2q-1}, v_{2q}) \in [N] \times [N]$, where the v_i's are not necessarily distinct, it holds that the statistical difference between $\mathrm{ans}_{G_1}(v_1, v_2), ..., \mathrm{ans}_{G_1}(v_{2q-1}, v_{2q})$ and $\mathrm{ans}_{G_2}(v_1, v_2), ..., \mathrm{ans}_{G_2}(v_{2q-1}, v_{2q})$ is $O(q^{3/2}\epsilon^2)$.*

Part 2 of Theorem 1.1 follows.

Proof: We consider a 1-1 correspondence, denoted ϕ, between the vertices of an N-vertex graph in $\mathcal{CC}_\epsilon \cup \mathcal{BCC}_\epsilon$ and triples in $[(4\epsilon)^{-1}] \times \{1, 2\} \times [2\epsilon \cdot N]$. Specifically, $\phi(v) = (i, j, w)$ indicates that v resides in the j^{th} "side" of the i^{th} part of the graph, and it is vertex number w in this set. That is, for a graph in $\mathcal{CC}_\epsilon$ the pair (i, j) indicates the j^{th} clique in the i^{th} pair of cliques, whereas for a graph in $\mathcal{BCC}_\epsilon$ the pair (i, j) indicates the j^{th} side in the i^{th} bi-cliques. Consequently, the answers provided by uniformly distributed $G_1 \in \mathcal{CC}_\epsilon$ and $G_2 \in \mathcal{BCC}_\epsilon$ can be emulated by the following two corresponding random processes.

1. The process A_1 selects uniformly a bijection $\phi : [N] \to [(4\epsilon)^{-1}] \times \{1, 2\} \times [2\epsilon \cdot N]$ and answers each query $(u, v) \in [N] \times [N]$ by 1 if and only if $\phi(u)$ and $\phi(v)$ agree on their first two coordinates (and differ on the third). That is, for $\phi(u) = (i_1, j_1, w_1)$ and $\phi(v) = (i_2, j_2, w_2)$, it holds that $A_1(u, v) = 1$ if and only if both $i_1 = i_2$ and $j_1 = j_2$ (and $w_1 \neq w_2$).
2. The process A_2 selects uniformly a bijection $\phi : [N] \to [(4\epsilon)^{-1}] \times \{1, 2\} \times [2\epsilon \cdot N]$ and answers each query $(u, v) \in [N] \times [N]$ by 1 if and only if $\phi(u) = (i, j, w_1)$ and $\phi(v) = (i, 3 - j, w_2)$. That is, for $\phi(u) = (i_1, j_1, w_1)$ and $\phi(v) = (i_2, j_2, w_2)$, it holds that $A_2(u, v) = 1$ if and only if $i_1 = i_2$ but $j_1 \neq j_2$.

Let us denote by $\phi'(v)$ (resp., $\phi''(v)$ and $\phi'''(v)$) the first (resp., second and third) coordinates of $\phi(v)$; that is, $\phi(v) = (\phi'(v), \phi''(v), \phi'''(v))$. Then, both processes answer the query (u, v) with 0 if $\phi'(u) \neq \phi'(v)$, and the difference between the processes is confined to the case that $\phi'(u) = \phi'(v)$. Specifically, conditioned on $\phi'(u) = \phi'(v)$ (and $\phi'''(u) \neq \phi'''(v)$), it holds that $A_1(u, v) = 1$ if and only if $\phi''(u) = \phi''(v)$, whereas $A_2(u, v) = 1$ if and only if $\phi''(u) \neq \phi''(v)$. However, since the (random) value of ϕ'' is not present at the answer, the forgoing difference may go unnoticed. The foregoing considerations apply to a single query, but things may change in case of several queries. For example, if $\phi'(u) = \phi'(v) = \phi'(w)$ then the answers to $(u, v), (v, w)$ and (w, v) will indicate whether we are getting answers from A_1 or from A_2 (since A_1 will answer positively on an odd number of these queries whereas A_2 will answer positively on an even number). In general, the event that allows distinguishing the two processes is an odd cycle of vertices that have the same ϕ' value. Minor differences may also be due to equal ϕ''' values, and so we also consider these in our "bad" event. For sake of simplicity, the bad event is defined more rigidly as follows, where the first condition represents the essential aspect and the second is a technicality.

Definition 4.1.1 *We say that ϕ is* bad (w.r.t the sequence $(v_1, v_2), ..., (v_{2q-1}, v_{2q}) \in [N] \times [N]$), *if one of the following two conditions hold:*

1. *For some $i \in [(4\epsilon)^{-1}]$, the subgraph $Q_i = (V_i, E_i)$, where $V_i = \{v_k : k \in [2q] \wedge \phi'(v) = i\}$ and $E_i = \{\{v_{2k-1}, v_{2k}\} : v_{2k-1}, v_{2k} \in V_i\}$, contains a simple cycle.*
2. *There exists $i \neq j \in [2q]$ such that $\phi'''(v_i) = \phi'''(v_j)$.*

Indeed, the query sequence $(v_1, v_2), ..., (v_{2q-1}, v_{2q})$ will be fixed throughout the rest of the proof, and so we shall omit it from our terminology.

Claim 4.1.2 *The probability that a uniformly distributed bijection ϕ is bad is at most*

$$2000 \cdot q^{3/2} \epsilon^2 + \frac{q^2}{2\epsilon N}$$

Proof: We start by upper-bounding the probability that the second event in Definition 4.1.1 holds. This event is the union of $\binom{2q}{2}$ sub-events, and each sub-event holds with probability $1/(4\epsilon \cdot N)$. Thus, we obtain a probability (upper) bound of $q^2/2\epsilon N$. As for the first event, for every $t \geq 3$, we upper-bound the probability that some Q_i contains a simple cycle of length t. We observe that the query graph $Q = (V_Q, E_Q)$, where $V_Q = \{v_k : k \in [2q]\}$ and $E_Q = \{\{v_{2k-1}, v_{2k}\} : k \in [q]\})$, contains at most $(2q)^{t/2}$ cycles of length t (cf. [A, Thm. 3]), whereas the probability that a specific simple t-cycle is contained in some Q_i is $(4\epsilon)^{t-1}$. Thus, the probability of the first event is upper-bounded by

$$\sum_{t\geq 3}(2q)^{t/2}\cdot(4\epsilon)^{t-1} < \sum_{t\geq 3}\left(\sqrt{2q}\cdot 4\cdot\epsilon^{(t-1)/t}\right)^t < \sum_{t\geq 3}\left(6\sqrt{q}\cdot\epsilon^{2/3}\right)^t,$$

which is upper-bounded by $2\cdot(6\sqrt{q}\cdot\epsilon^{2/3})^3 < 500q^{3/2}\epsilon^2$, provided $6\sqrt{q}\cdot\epsilon^{2/3} < 1/2$ (and the claim hold trivially otherwise). $\square$

Claim 4.1.3 *Conditioned on the bijection ϕ not being bad, the sequences $(A_1(v_1, v_2), ..., A_1(v_{2q-1}, v_{2q}))$ and $(A_2(v_1, v_2), ..., A_2(v_{2q-1}, v_{2q}))$ are identically distributed.*

Proof: Noting that Definition 4.1.1 only refers to ϕ' and ϕ''', we fixed any choice of ϕ' and ϕ''' that yields a good ϕ and consider the residual random choice of ϕ''. Referring to the foregoing subgraphs Q_i's, recall that pairs with endpoints in different Q_i's are answered by 0 in both processes. Note that (by the second condition in Definition 4.1.1) the hypothesis implies that ϕ''' assigns different values to the different vertices in $\{v_k : k \in [2q]\}$, and it follows that ϕ'' assigns these vertices values that are uniformly and independently distributed in $\{1, 2\}$. Now, using the first condition in Definition 4.1.1, the hypothesis implies that each Q_i is a forest, which implies that (in each of the two processes) the answer assigned to each edge in Q_i is independent of the answer given to other edges of Q_i. That is, we assert that (in each of the two processes) the edges of each forest $Q_i = (V_i, E_i)$ are assigned a sequence of answers that is uniformly distributed in $\{0, 1\}^{|E_i|}$. To formally prove this assertion, consider the constraints on the ϕ''-values (of V_i) that arise from any possible sequence of answers. These constraints form a system of $|E_i|$ linear equations over $GF(2)$ with variables corresponding to the possible ϕ''-values and constant terms encoding possible equality and inequality constraints.[10] Note that the (coefficients of the) linear systems are not affected by the identity of the process, which does effect the free terms. Furthermore, this linear system is of full rank; and thus, for each of the two processes and each sequence of answers, the corresponding system has $2^{|V_i|-|E_i|} = 2$ solutions (i.e., possible assignments to ϕ'' restricted to V_i). Thus, in each of the two processes, each query is answered by the value 1 with probability exactly 1/2, independently of the answers to all other queries. The claim follows. $\square$

Combining Claims 4.1.2 and 4.1.3, it follows that the statistical distance between the sequences $(A_1(v_1, v_2), ..., A_1(v_{2q-1}, v_{2q}))$ and $(A_2(v_1, v_2), ..., A_2(v_{2q-1}, v_{2q}))$ is at most $O(q^{3/2}\epsilon^2 + q^2(\epsilon N)^{-1})$, and the lemma follows for sufficiently large N. $\blacksquare$

4.2 A Matching Upper-Bound

In this section we establish Part 3 of Theorem 1.1. We mention that this improves over the $\widetilde{O}(\epsilon^{-2})$ bound of [AS, Thm. 2] (which is based on inspecting the subgraph induced by a random set of $O(\epsilon^{-1}\log(1/\epsilon))$ vertices).

Algorithm 4.2 (non-adaptive test for $\mathcal{CC}$): *On input N and ϵ and oracle access to a graph $G = ([N], E)$, the tester sets $\ell = \log_2(1/\epsilon)$ and proceeds as follows.*

[10] The condition $A_1(u, w) = 1$ iff $\phi''(u) = \phi''(v)$ is encoded by $\phi''(u) + \phi''(v) = A_1(u, w) + 1$, whereas the condition $A_2(u, w) = 1$ iff $\phi''(u) \neq \phi''(v)$ is encoded by $\phi''(u) + \phi''(v) = A_2(u, w)$.

1. *The tester selects a random sample of* $s \stackrel{\rm def}{=} \Theta(\epsilon^{-2/3})$ *vertices, denoted* S, *and examines all vertex pairs* (in $S \times S$).

2. *For* $i = 1, ..., (2\ell/3) + \Theta(1)$, *the tester selects uniformly a subset* $S_i \subseteq S$ *of cardinality* $s_i \stackrel{\rm def}{=} \Theta(2^i)$ *and a sample of* $\widetilde{\Theta}(\epsilon^{-1})/s_i$ *vertices, denoted* R_i, *and examines all the vertex pairs in* $S_i \times R_i$.

3. *The tester accepts if and only if its view of the graph as obtained in Steps 1-2 is consistent with some graph in* $\mathcal{CC}$. *Namely, let* $g' : ((S \times S) \cup \bigcup_{i=1}^{\ell'}(S_i \times R_i)) \to \{0,1\}$ *be the function determined by the answers obtained in Steps 1-2. Then the tester accepts if and only if* g' *can be extended to a function over* $S' \times S'$, *where* $S' = S \cup \bigcup_{i=1}^{\ell'} R_i$, *that represents a graph in* $\mathcal{CC}$.

The query complexity of Algorithm 4.2 is dominated by Step 1, which uses $O(\epsilon^{-2/3})^2 = O(\epsilon^{-4/3})$ queries. Clearly, this algorithm accepts (with probability 1) any graph that is in $\mathcal{CC}$. It remains to analyze its behavior on graphs that are ϵ-far from $\mathcal{CC}$.

Lemma 4.3 *If* $G = ([N], E)$ *is* ϵ*-far from* $\mathcal{CC}$, *then on input* N, ϵ *and oracle access to* G, *Algorithm 4.2 rejects with probability at least* $2/3$.

Part 3 of Theorem 1.1 follows.

Proof: We say that a triple (v, u, w) of vertices (resp., a 3-set $\{v, u, w\} \subset [N]$) is a witness (for rejection) *if the subgraph* of G *induced by* $\{v, u, w\}$ *contains exactly two edges.* Indeed, Algorithm 4.2 rejects if (and only if), for some witness (v, u, w), the algorithm has made all three relevant queries (i.e., the queries (v, u), (u, w), and (w, v)).[11] A sufficient condition for this to happen is that either $\{v, u, w\} \subset S$ or for some i both $|\{v, u, w\} \cap S_i| = 2$ and $|\{v, u, w\} \cap R_i| = 1$ hold. Thus, we say that a witness is effective with respect to the said samples (i.e., S and the R_i's) if the foregoing sufficient condition holds. We shall show that, with probability at least $2/3$, the samples contain an effective witness.

Let $G' = (V, E')$ be a graph in $\mathcal{CC}$ that is closest to $G = (V, E)$, and let $(V_1, ..., V_t)$ be its partition into cliques. For the sake of simplicity, we shall refer to the V_i's as cliques, even though they are not (necessarily) cliques in G, and we shall refer to the partition $(V_1, ..., V_t)$ as the *best possible partition* for G. Two main observations regarding this partition follow.

Observation 1: For every $i \in [t]$ and every $S \subseteq V_i$, it holds that $|E \cap (S \times (V_i \setminus S))| \geq |S \times (V_i \setminus S)|/2$, since otherwise replacing the clique V_i by two cliques, S and $V_i \setminus S$ yields a better partition for G.

Observation 2: For every $i \neq j \in [t]$, it holds that $|E \cap (V_i \times V_j)| \leq |V_i \times V_j|/2$, since otherwise replacing the two cliques V_i and V_j by a single clique $V_i \cup V_j$ yields a better partition for G.

Now, since G is ϵ-far from $\mathcal{CC}$, either G misses $\frac{\epsilon}{2} \cdot N^2$ edges within these V_i's or it has $\frac{\epsilon}{2} \cdot N^2$ superfluous edges between distinct V_i's. We show that in either case, with high constant probability, the samples produced by Algorithm 4.2 contain an effective witness.

The pivot of the analysis is relating the fraction of bad vertex pairs (i.e., either missing "internal" edges or superfluous "external" edges) to the fraction of witnesses. Specifically, we shall show that the existence of $\frac{\epsilon}{2} \cdot N^2$ missing internal edges (resp., $\frac{\epsilon}{2} \cdot N^2$ superfluous "external" edges) implies

[11] We note that only the (easy to establish) sufficiency of the foregoing rejection condition is used in the analysis.

the existence of $\Omega(\epsilon^2 N^3)$ witnesses. Furthermore, using additional features of the structure of the set of witnesses, we shall show that with high probability the random sample (as produced by Algorithm 4.2) contains an effective witness. Specifically, these additional features, which are established in the elaborate parts of Claims 4.3.1 and 4.3.2, are instrumental to the detection of a witness (as argued in Claim 4.3.3).

To facilitate the exposition, for every two sets $A, B \subset [N]$, we let $E(A,B)$ denote the set of edges with one endpoint in A and another endpoint in B (i.e., $E(A,B) \stackrel{\text{def}}{=} E \cap (A \times B)$). For each vertex v and $j \in [t]$, let

$$\Gamma_j(v) \stackrel{\text{def}}{=} V_j \cap \Gamma(v) = \{u \in V_j : (u,v) \in E\}$$

and

$$\overline{\Gamma}_j(v) \stackrel{\text{def}}{=} V_j \setminus (\Gamma(v) \cup \{v\}) = \{u \in (V_j \setminus \{v\}) : (u,v) \notin E\}.$$

If $v \in V_i$, then we use the shorthand: $\overline{\Gamma}(v) = \overline{\Gamma}_i(v)$. Indeed, $\overline{\Gamma}(v)$ corresponds to the set of internal edges that are missed by vertex v.

Claim 4.3.1 (using missing internal edges):

Basic claim: *For every vertex v, the number of witnesses that contain v is $\Omega(|\overline{\Gamma}(v)|^2)$.*

Elaborate claim: *For every* (possibly empty)[12] *set F of* "forbidden" (non-adjacent) *vertex-pairs, the following holds:*

1. *For every $v \in [N]$ there exists a set $W_v \subseteq \overline{\Gamma}(v) \setminus \{u : (v,u) \in F\}$ such that*

$$\sum_{v \in [N]} |W_v| > \left(\sum_{v \in [N]} \frac{|\overline{\Gamma}(v)|}{4} \right) - 2 \cdot |F|$$

   and for every $u \in W_v$ there exists a set $W_{v,u} \subseteq (\Gamma(v) \cap \Gamma(u))$ such that

$$\sum_{u \in W_v} |W_{v,u}| \geq |W_v|^2/4.$$

   Moreover, if $F = \emptyset$ then for every v it holds that $|W_v| \geq |\overline{\Gamma}(v)|/4$.

   (Indeed, each triple (u,v,w) such that $u \in W_v$ and $w \in W_{v,u}$ constitutes a witness, because $\{u,v\} \notin E$ whereas $w \in \Gamma(v) \cap \Gamma(u)$; see illustration in Figure 1.)

2. *For the sets W_v and $W_{v,u}$ as in Part 1 of the claim, letting $U_w^{(2)} \stackrel{\text{def}}{=} \{(v,u) : w \in W_{v,u}\}$ it holds that if each set W_v has cardinality at most $\epsilon^{2/3}N/2$ then each $U_w^{(2)}$ has cardinality at most $\epsilon^{4/3}N^2$.*

It follows that the total number of witnesses is $\Omega(\sum_{v \in [N]} |\overline{\Gamma}(v)|^2)$. In particular, if the number of missing internal edges is at least $\frac{\epsilon}{2} \cdot N^2$ (i.e., $\sum_{v \in [N]} |\overline{\Gamma}(v)| \geq \epsilon \cdot N^2$), then the total number of witnesses is at least $N \cdot \Omega((\epsilon N)^2) = \Omega(\epsilon^2 \cdot N^3)$.

Proof: Using Observation 1, we note that for any choice of $i \in [t]$ and for every $v \in V_i$ it holds that

$$|\overline{\Gamma}(v)| = |V_i \setminus \{v\}| - |E(\{v\}, V_i \setminus \{v\})| \leq \frac{|V_i| - 1}{2} \leq |\Gamma_i(v)| \tag{3}$$

[12] Indeed, in first reading, the reader is encouraged to think of the case $F = \emptyset$. In fact, this case is one of the two cases that will be actually used in the sequel.

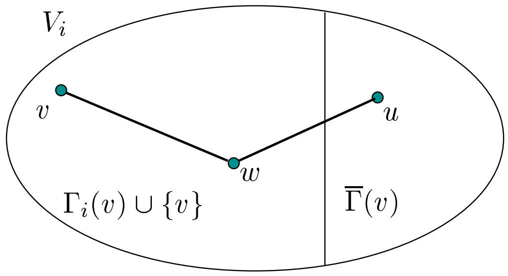

Figure 1: An Illustration for the proof of Claim 4.3.1.

and

$$|E(\overline{\Gamma}(v), \Gamma_i(v)| \;=\; |E(\overline{\Gamma}(v), \Gamma_i(v) \cup \{v\})| \;>\; \frac{1}{2}|\overline{\Gamma}(v)| \cdot |\Gamma_i(v)| \,. \tag{4}$$

Letting $T_v = \{(v,u,w) : (u,w) \in \overline{\Gamma}(v) \times \Gamma_i(v)\}$, it follows that at least half of the triples (v,u,w) in T_v are witnesses (i.e., $(u,w) \in E$, $(u,v) \notin E$, and $(w,v) \in E$), whereas $|T_v| \geq |\overline{\Gamma}(v)|^2$. This establishes the basic claim.

Let us first establish the elaborate claim for the special case of $F = \emptyset$. In this case, for every $v \in V_i$, we consider the set

$$W_v \stackrel{\rm def}{=} \{u \in \overline{\Gamma}(v) : |E(\{u\}, \Gamma_i(v)| \geq |\Gamma_i(v)|/4\} \,. \tag{5}$$

By Eq. (4), $\sum_{u \in \overline{\Gamma}(v)} |E(\{u\}, \Gamma_i(v))| \geq |\overline{\Gamma}(v)| \cdot |\Gamma_i(v)|/2$. It follows that $|W_v| \geq |\overline{\Gamma}(v)|/4$. We note that (by Eq. (5)), for every $u \in W_v$, it holds that $|\Gamma_i(v) \cap \Gamma(u)| \geq |\Gamma_i(v)|/4 \geq |W_v|/4$. Next, for every $u \in W_v$, let $W_{v,u}$ be an arbitrary subset of $|W_v|/4$ elements in $\Gamma_i(v) \cap \Gamma(u)$. Note that, indeed $W_v \subseteq \overline{\Gamma}(v)$ and for every $u \in W_v$ it holds that $W_{v,u} \subseteq \Gamma(v) \cap \Gamma(u)$. Recalling that $|W_v| \geq |\overline{\Gamma}(v)|/4$ and $|W_{v,u}| = |W_v|/4$, Part 1 follows.

To establish Part 2, we first note that if we select $W_{v,u}$ uniformly among all $|W_v|/4$-subsets of $\Gamma_i(v) \cap \Gamma(u)$, then, for any $w \in V_i$, the expected size of $U_w^{(2)}$ is upper-bounded by

$$\sum_{v \in V_i} \sum_{u \in W_v} \frac{|W_v|/4}{|\Gamma_i(v) \cap \Gamma(u)|} \;\leq\; \sum_{v \in V_i} \sum_{u \in W_v} \frac{|W_v|/4}{|V_i|/8} \;=\; \frac{2}{|V_i|} \cdot \sum_{v \in V_i} |W_v|^2$$

where the inequality uses $|\Gamma_i(v) \cap \Gamma(u)| \geq |\Gamma_i(v)|/4 \geq |V_i|/8$. Thus, if $\frac{2}{|V_i|} \cdot \sum_{v \in V_i} |W_v|^2 \leq \epsilon^{4/3} N^2/2$ then, with overwhelmingly high probability, it holds that $|U_w^{(2)}| \leq \epsilon^{4/3} N^2$. Picking the sets (i.e., the $W_{v,u}$'s) such that none of the negligible probability events (associated with $w \in V_i$) occurs, we infer that $|U_w^{(2)}| > \epsilon^{4/3} N^2$ implies that $\sum_{v \in V_i} |W_v|^2 > \epsilon^{4/3} N^2 |V_i|/4$ (which implies the existence of v such that $|W_v| > \epsilon^{2/3} N/2$). Part 2 follows.

Note that so far we have established the (elaborate) claim for the special case of $F = \emptyset$. We now establish the general case by reduction to the former special case. We first modify the sets W_v, by omitting from each W_v each vertex u such that $\{v,u\} \in F$. This modification decreases $\sum_v |W_v|$ by at most $2|F|$. Next, we modify the sets $W_{v,u}$ by omitting from each $W_{v,u}$ a few elements, selected at random, such that $|W_{v,u}| = |W_v|/4$ holds (for the modified sets). Clearly, Part 1 holds for the modified sets. To see that Part 2 holds too, we note that the foregoing argument only relies on the fact that $W_{v,u}$ is a random $(|W_v|/4)$-size subset of $\Gamma_i(v) \cap \Gamma(u)$. The claim follows. $\square$

Another piece of notation. For every $i \in [t]$ and every $v \in V_i$, let

$$\Gamma'(v) \stackrel{\text{def}}{=} \Gamma(v) \setminus V_i$$

denote the set of vertices outside of V_i that have a superfluous edge to v. That is, $\Gamma'(v) = \bigcup_{j \neq i} \Gamma_j(v)$.

Claim 4.3.2 (using superfluous external edges):

Basic claim: *For every vertex v, the number of witnesses that contain v is $\Omega(|\Gamma'(v)|^2)$.*

Elaborate claim: *If $\sum_{v \in [N]} |\Gamma'(v)| > 500 \cdot \sum_{v \in [N]} |\overline{\Gamma}(v)|$, then there exist constants $c_1, \ldots, c_4$ for which the following holds:*

1. *For every $v \in [N]$ there exists a set $W_v \subseteq \Gamma'(v)$ such that letting $V' = \{v : |W_v| \geq |\Gamma'(v)|/c_1\}$ it holds that*

$$\sum_{v \in V'} |\Gamma'(v)| \geq \frac{3}{4} \sum_{v \in [N]} |\Gamma'(v)| \,. \tag{6}$$

   In addition, for every $u \in W_v$ there exists a set $W_{v,u}$, which is either a subset of $\Gamma(v) \setminus \Gamma(u)$ or a subset of $\Gamma(u) \setminus \Gamma(v)$, such that $|W_{v,u}| \geq |W_v|/c_2$.
   (Indeed, each (v, u, w) such that $u \in W_v$ and $w \in W_{v,u}$ constitutes a witness.)
2. *For the sets $W_{v,u}$ as in Part 1 of the claim, let $U_w^{(2)} \stackrel{\text{def}}{=} \{(v, u) : w \in W_{v,u}\}$. If for every v it holds that $|\Gamma'(v)| \leq \epsilon^{2/3} N/2$ then each $U_w^{(2)}$ has cardinality at most $10\epsilon^{4/3} N^2$.*
3. *Let F be any set of* "forbidden" *vertex-pairs in $\bigcup_{i \neq j} E(V_i, V_j)$, and for a vertex v let $F(v) \stackrel{\text{def}}{=} \{u : (v, u) \in F\}$. Then, for each vertex v, there exist modified subsets W_v and $W_{v,u}$ (for every $u \in W_v$) that satisfy the following modified versions of Parts 1 and 2:*
   - *For Part 1 it holds that $W_v \subseteq \Gamma'(v) \setminus F(v)$, and Eq.* (6) *is replaced by*

$$\sum_{v \in [N]} |W_v| \; > \; \frac{1}{c_3} \left(\sum_{v \in [N]} |\Gamma'(v)| \right) \; - \; c_4 \cdot |F| \,. \tag{7}$$

     The other features of the subsets W_v and $W_{v,u}$ hold as stated in Part 1.
   - *For Part 2 we have that if for every v it holds that $|\Gamma'(v) \setminus F(v)| \leq \epsilon^{2/3} N/2$ then each modified $U_w^{(2)}$* (i.e., $U_w^{(2)} \stackrel{\text{def}}{=} \{(v, u) : w \in W_{v,u}\}$) *has cardinality at most $10\epsilon^{4/3} N^2$.*

It follows that the total number of witnesses is $\Omega(\sum_{v \in [N]} |\Gamma'(v)|^2)$. In particular, if the number of superfluous external edges is at least $\frac{\epsilon}{2} \cdot N^2$ (i.e., $\sum_{v \in [N]} |\Gamma'(v)| \geq \epsilon \cdot N^2$), then the total number of witnesses is at least $N \cdot \Omega((\epsilon N)^2) = \Omega(\epsilon^2 \cdot N^3)$.

Proof: We first prove Parts 1 and 2, and later present the modifications required for Part 3. The claim is proved by a (rather tedious) case analysis. In all but one of the cases, the basic claim (i.e., for every vertex v, the number of witnesses that contain v is $\Omega(|\Gamma'(v)|^2)$) follows from the elaborate claim, and so in those cases it suffices to prove the latter. In the exceptional case, the basic claim follows by invoking Claim 4.3.1.

Each case deals with a different subset of vertices of V. With the exception of the aforementioned case, Part 1 is proved by presenting, for every relevant vertex v (i.e., v that satisfies the case hypothesis), a subset $W_v \subseteq \Gamma'(v)$ of size at least $\Gamma'(v)/c_1$ and adequate sets $W_{v,u}$ for each $u \in W_v$.

Furthermore, it will be shown that the vertices covered by these (non-exceptional cases) account for at least three fourths of the sum $\sum_v |\Gamma'(v)|$.

In order to prove Part 2, for each of the foregoing cases, we consider the restriction of $U_w^{(2)}$ to pairs (v, u) such that v obeys the case hypothesis. We show that if $|\Gamma'(v)| \leq \epsilon^{2/3}N/2$ for every such v, then the total contribution to $U_w^{(2)}$ of the corresponding pairs (v, u) is at most $\epsilon^{4/3}N^2$. Since there are less than ten cases, Part 2 follows.

In the following analysis we consider possible cases that may apply to a generic vertex v. However, we actually consider the set of all vertices that satisfy the hypothesis of each of these cases. Hence, when we say that Part 1 (resp., Part 2) is established for the vertices that satisfy a particular case hypothesis, we mean that the condition is established in the sense described in the foregoing discussion. We now turn to the actual case analysis.

Case 1: Much of $\Gamma'(v)$ is contained in a single V_j; that is, there exists an index j such that $|\Gamma_j(v)| > |\Gamma'(v)|/10$. Fixing such an index j, we distinguish two subcases regarding the fraction of V_j that is not covered by $\Gamma'(v)$ (i.e., the relative density of $\overline{\Gamma}_j(v)$ in V_j).

Case 1.1: $|\overline{\Gamma}_j(v)| \geq |V_j|/10$. In this case, we let W_v be a subset of the neighbors that v has in V_j, that is, a subset of $\Gamma_j(v)$. For each $u \in W_v$ we let $W_{v,u}$ be a subset of the non-neighbors of v in V_j that are neighbors of w, that is, a subset of $\overline{\Gamma}_j(v) \cap \Gamma_j(u)$. Thus, for every $u \in W_v$ and $w \in W_{v,u}$, the triple (v, u, w) is a witness. For an illustration, see Figure 2. Combining this case hypothesis (which asserts that v has many non-neighbors in V_j) with Observation 1 (which guarantees many edges between neighbors and non-neighbors of v in V_j), we obtain many (i.e., $\Omega(|\Gamma'(v)|^2)$) such witnesses, and the basic claim follows.

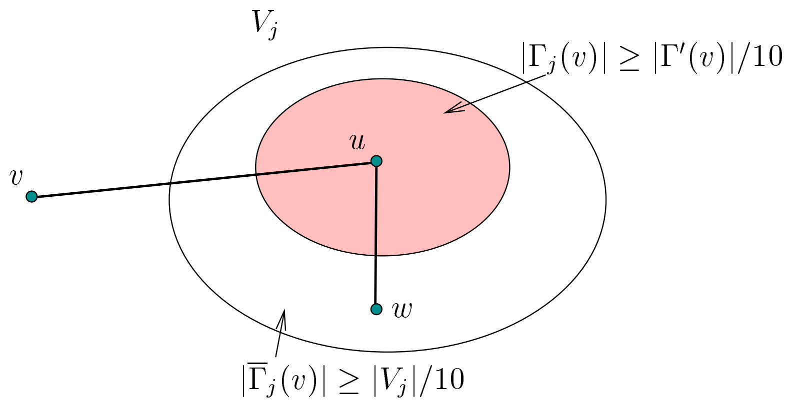

Figure 2: An Illustration for the proof of Claim 4.3.2, Case 1.1.

In order to actually prove Parts 1 and 2, we now provide a more detailed description of the choice of W_v and $W_{v,u}$. Let the subset of vertices for which the case (1.1) hypothesis holds be denoted by $V^{1.1}$. For each vertex $v \in V^{1.1}$, let $\xi(v) \stackrel{\rm def}{=} j$ if j is the smallest integer such that $|\Gamma_j(v)| > |\Gamma'(v)|/10$. Next, we define the set

$$W_v \stackrel{\rm def}{=} \{u \in \Gamma_{\xi(v)}(v) : |\Gamma(u) \cap (\overline{\Gamma}_{\xi(v)}(v))| \geq |\overline{\Gamma}_{\xi(v)}(v)|/4\},$$

and note that (by the case hypothesis) for every $u \in W_v$ it holds that $|\Gamma(u) \cap (\overline{\Gamma}_{\xi(v)}(v))| \geq |V_{\xi(v)}|/40$. By Observation 1, $|E(\Gamma_{\xi(v)}(v), \overline{\Gamma}_{\xi(v)}(v))| \geq |\Gamma_{\xi(v)}(v)| \cdot |\overline{\Gamma}_{\xi(v)}(v)|/2$. Noting that

$|E(\Gamma_{\xi(v)}(v),\overline{\Gamma}_{\xi(v)}(v))| = \sum_{u\in\Gamma_{\xi(v)}(v)} |\Gamma(u)\cap(\overline{\Gamma}_{\xi(v)}(v))|$ and referring to the definition of W_v, it follows that $|W_v| \geq |\Gamma_{\xi(v)}(v)|/4 \geq |\Gamma'(v)|/40$.

Now, for every $u \in W_v$, let $W_{v,u}$ be a random subset of $|W_v|/40$ elements in $\overline{\Gamma}_{\xi(v)}(v)\cap\Gamma(u)$, while recalling that the latter set has size at least $|\overline{\Gamma}_{\xi(v)}(v)|/4 \geq |V_{\xi(v)}|/40$. Observe that indeed, for every $u \in W_v$ and $w \in W_{v,u}$, it holds that $W_v \subseteq \Gamma'(v)$ and $W_{v,u} \subseteq \Gamma(u)\setminus\Gamma(v)$. (We note that for every $w \in W_{v,u}$ it holds that $w \notin \Gamma(v)$ and $w \in \Gamma(u)\setminus\Gamma'(u)$ (since both $u \in V_{\xi(v)}$ and $W_{v,u} \subseteq V_{\xi(v)}$).) Part 1 is thus established for this case (for any $v \in V^{1.1}$).

To establish Part 2, we first note that, for any $j \in [t]$ and $w \in V_j$, the expected size of $U_w^{(2)}$ is upper-bounded by

$$\sum_{v\in V^{1.1}:\xi(v)=j}\ \sum_{u\in W_v} \frac{|W_v|/40}{|\overline{\Gamma}_j(v)\cap\Gamma(u)|} \ \leq\ \frac{1}{|V_j|}\cdot\sum_{v\in V^{1.1}:\xi(v)=j}|W_v|^2$$

where the inequality uses $|\overline{\Gamma}_j(v)\cap\Gamma(u)| \geq |V_j|/40$. As in the proof of Claim 4.3.1, it is possible to choose the subsets $W_{v,u}$ so that the sizes of the sets $U_w^{(2)}$ are not much larger than (the upper bounds on the value of) their expected sizes. It follows that if some $w \in V_j$ satisfies $|U_w^{(2)}| > \epsilon^{4/3}N^2$, then $\sum_{v\in V^{1.1}:\xi(v)=j}|W_v|^2 > \epsilon^{4/3}N^2|V_j|/2$. We now consider two cases. In the easy case there exists a vertex v for which $\xi(v) = j$ and such that $|W_v| > \epsilon^{2/3}N/2$, and Part 2 follows (since $W_v \subseteq \Gamma'(v)$). Otherwise, letting $V' = \{v \in V^{1.1} : \xi(v) = j\}$, we note that

$$|E(V',V_j)| \ \geq\ \sum_{v\in V'}|W_v| \ \geq\ \sum_{v\in V'}\frac{|W_v|^2}{\epsilon^{2/3}N/2} \ >\ |V_j|\cdot\epsilon^{2/3}N \tag{8}$$

and it follows that there exists a vertex $u \in V_j$ such that $|\Gamma'(u)| \geq |\Gamma(u)\cap V'| > \epsilon^{2/3}N$. Thus, Part 2 follows in this case.

Case 1.2: $|\overline{\Gamma}_j(v)| \leq |V_j|/10$ (i.e., $|\Gamma_j(v)| \geq 0.9|V_j|$). We first note that $|\Gamma_i(v)| \geq 0.8|\Gamma_j(v)|$, because otherwise we would obtain a better partition by moving the vertex v from V_i to V_j (since the gain from such a move is at least $(|\Gamma_j(v)| - |\overline{\Gamma}_j(v)|) - |\Gamma_i(v)|$, whereas $|\Gamma_j(v)| - |\overline{\Gamma}_j(v)| \geq 0.8|V_j| \geq 0.8|\Gamma_j(v)|$). We consider two subcases regarding the cardinality of the set $\Gamma_i(v)$:

1. If $|\Gamma_i(v)| \geq 0.9\cdot|V_i|$, then we let W_v be a subset of $\Gamma_j(v)$, and for each $u \in W_v$, we let $W_{v,u}$ be a subset of $\Gamma_i(v)\setminus\Gamma(u)$. Thus each triple (v,u,w) where $u \in W_v$ and $w \in W_{v,u}$ is a witness. For an illustration, see Figure 3. Combining the case hypotheses (which asserts that $V_j \times V_i$ is essentially covered by $\Gamma_j(v)\times\Gamma_i(v)$) with Observation 2 (which guarantees many non-edges in $V_j \times V_i$), we obtain $\Omega(|\Gamma'(v)|^2)$ such witnesses. Details follow.
   Let the subset of vertices for which the case hypothesis holds be denoted by $V^{1.2}$, and for each $v \in V^{1.2}$ define $\xi(v)$ as in Case 1.1. Let

   $$W_v \stackrel{\rm def}{=} \{u\in\Gamma_j(v) : |\Gamma_i(v)\setminus\Gamma(u)| \geq |\Gamma_i(v)|/10\}\,.$$

   Note that for any $u \in W_v$ it holds that $|\Gamma_i(v)\setminus\Gamma(u)| \geq 0.1|\Gamma_i(v)| \geq 0.08|\Gamma_j(v)|$. Using Observation 2 we have that

   $$\begin{aligned}|E(\Gamma_j(v),\Gamma_i(v))| &\leq |E(V_j,V_i)|\\ &\leq \frac{1}{2}\cdot|V_j|\cdot|V_i|\end{aligned}$$

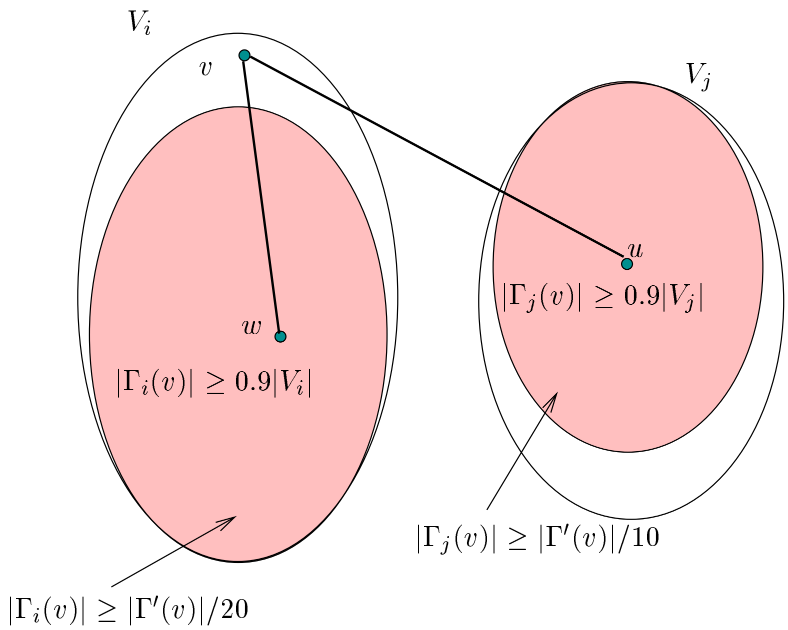

Figure 3: An Illustration for the proof of Claim 4.3.2, 1st subcase of Case 1.2.

$$\begin{aligned} &\leq \frac{1}{2} \cdot \frac{|\Gamma_j(v)|}{0.9} \cdot \frac{|\Gamma_i(v)|}{0.9} \\ &< 0.7 \cdot |\Gamma_j(v)| \cdot |\Gamma_i(v)| \ . \end{aligned}$$

Hence there are at least $0.3 \cdot |\Gamma_j(v)| \cdot |\Gamma_i(v)|$ pairs (u, w) where $u \in \Gamma_j(v)$ and $w \in \Gamma_i(v)$ such that $w \notin \Gamma(u)$. It follows that $|W_v| > |\Gamma_j(v)|/5$, where by the hypothesis of Case 1 this value is greater than $|\Gamma'(v)|/50$.

Next, recalling that for any $u \in W_v$ it holds that $|\Gamma_i(v) \setminus \Gamma(u)| \geq 0.08|\Gamma_j(v)|$, we let $W_{v,u}$ be a $0.08|W_v|$-size random subset of $\Gamma_i(v) \setminus \Gamma(u) \subseteq \Gamma(v) \setminus \Gamma(u)$, and note that indeed for every $u \in W_v$ and $w \in W_{v,u}$ it holds that $u, w \in \Gamma(v)$ and $(u, w) \notin E$. Thus, Part 1 follows in this case. (We note that for every $w \in W_{v,u}$ it holds that $w \notin \Gamma(u)$ and $w \in \Gamma(v) \setminus \Gamma'(v)$ (since $v, w \in V_i$).)

As for Part 2, we first note that for every $w \in V_i$ the expected size of $U_w^{(2)}$ (in this case) is upper-bounded by

$$\sum_{v \in V_i} \sum_{u \in W_v} \frac{0.08|W_v|}{|\Gamma_i(v) \setminus \Gamma(u)|} \leq \frac{0.08}{0.09|V_i|} \cdot \sum_{v \in V_i} |W_v|^2$$

where the inequality uses $|\Gamma_i(v) \setminus \Gamma(u)| \geq 0.1|\Gamma_i(v)| \geq 0.09|V_i|$. Again, we may select the sets $W_{v,u}$ such that for each $w \in V_i$ it holds that $|U_w^{(2)}| < \sum_{v \in V_i} |W_v|^2/|V_i|$. Thus, if some $w \in V_i$ satisfies $|U_w^{(2)}| > \epsilon^{4/3}N^2$, then $\sum_{v \in V_i} |W_v|^2 > \epsilon^{4/3}N^2|V_i|$. It follows that there exists a vertex $v \in V_i$ such that $|W_v| > \epsilon^{2/3}N$, and Part 2 follows.

2. If $|\Gamma_i(v)| \leq 0.9 \cdot |V_i|$, then we proceed somewhat differently than in the other cases (this is the exceptional case mentioned at the preamble of the proof). Recall that $\overline{\Gamma}(v) = \overline{\Gamma}_i(v) = V_i \setminus \Gamma(v)$, and so $|\overline{\Gamma}(v)| \geq 0.1 \cdot |V_i| \geq 0.008 \cdot |\Gamma'(v)|$ (because $|V_i| \geq |\Gamma_i(v)| \geq 0.8|\Gamma_j(v)|$ and $|\Gamma_j(v) \geq |\Gamma'(v)|/10$). For the basic claim, we invoke Claim 4.3.1,

translating the lower-bound in terms of $|\overline{\Gamma}(v)|$ (provided by Claim 4.3.1) into a lower-bound in terms of $|\Gamma'(v)|$. For the elaborate claim, we set $W_v = \emptyset$ for every v as in the case hypothesis. Thus we trivially have that $|W_{v,u}| \geq |W_v|/c_2$ for every $u \in W_v$, and Part 2 of the claim holds trivially as well. Finally, we use the premise of the claim that $\sum_{v\in[N]} |\Gamma'(v)| > 500 \sum_{v\in[N]} |\overline{\Gamma}(v)|$ to infer that the current subcase (in which $|\Gamma'(v)| \leq 125|\overline{\Gamma}(v)|$) may account for less than one fourth of the sum $\sum_{v\in[N]} |\Gamma'(v)|$.

This completes the treatment of the current case (i.e., Case 1.2), which in turn completes the treatment of Case 1. (We thus proceed to the following complementary Case 2.)

Case 2: No single V_j contains much of $\Gamma'(v)$; that is, for every j it holds that $|\Gamma_j(v)| \leq |\Gamma'(v)|/10$. As in Case 1, we consider two subcases regarding the relative part of each V_j covered by $\Gamma'(v)$, but in the current case we consider a partition of the set $J \stackrel{\rm def}{=} \{j : |\Gamma_j(v)| \geq 1\}$ and distinguish cases regarding the intersection of $\Gamma'(v)$ with the sets V_j in each part.[13] Specifically, we let $J' \stackrel{\rm def}{=} \{j : |\Gamma_j(v)| > 0.9|V_j|\}$, and consider the following two subcases.

Case 2.1: $\sum_{j\in J'} |\Gamma_j(v)| \geq 0.5 \cdot |\Gamma'(v)|$. In this case J' has cardinality at least five (since $\sum_{j\in J'} |\Gamma_j(v)| \geq 0.5 \cdot |\Gamma'(v)|$ and $|\Gamma_j(v)| \leq 0.1 \cdot |\Gamma'(v)|$ for every j). Let $C_v = \bigcup_{j\in J'} \Gamma_j(v)$ (note that the vertices in C_v belong to several cliques V_j). In this case we let W_v be a subset of C_v, and for each $u \in C_v$ we let $W_{v,u}$ be a subset of $C_v \setminus \Gamma(u)$. We shall show that the case hypothesis implies that there are many missing edges between pairs of vertices in C_v. Intuitively, this holds because C_v essentially covers $\bigcup_{j\in J'} V_j$, whereas (by Observation 2) for any $j_1 \neq j_2$ there are many non-edges in $V_{j_1} \times V_{j_2}$. This ensures that we have many witnesses of the form (v,u,w), where $u \in W_v$ and $w \in W_{v,u}$. Details follow.

For every $j_1 \neq j_2 \in J'$, by Observation 2 (and since $|\Gamma_j(v)| > 0.9|V_j|$ for every $j \in J'$), it holds that

$$|E(\Gamma_{j_1}(v), \Gamma_{j_2}(v))| \leq \frac{1}{2} \cdot |V_{j_1}| \cdot |V_{j_2}| < 0.7 \cdot |\Gamma_{j_1}(v)| \cdot |\Gamma_{j_2}(v)| \,.$$

Letting $M \stackrel{\rm def}{=} \sum_{j_1\neq j_2\in J'} |(\Gamma_{j_1}(v) \times \Gamma_{j_2}(v)) \setminus E|$, we first observe that

$$\begin{aligned}
M &= \sum_{j_1\neq j_2\in J'} (|\Gamma_{j_1}(v)| \cdot |\Gamma_{j_2}(v)| - |E(\Gamma_{j_1}(v), \Gamma_{j_2}(v))|) \\
&\geq \sum_{j_1\neq j_2\in J'} (1-0.7) \cdot |\Gamma_{j_1}(v)| \cdot |\Gamma_{j_2}(v)| \\
&= 0.3 \cdot \left(\left(\sum_{j\in J'} |\Gamma_j(v)| \right)^2 - \sum_{j\in J'} |\Gamma_j(v)|^2 \right) \\
&\geq 0.3 \cdot \left((0.5 \cdot |\Gamma'(v)|)^2 - 0.1 \cdot |\Gamma'(v)|^2 \right),
\end{aligned}$$

where the last inequality uses the hypotheses of Cases 2 and 2.1. Therefore, $|(C_v \times C_v) \setminus E| \geq M > 0.04 \cdot |\Gamma'(v)|^2$.

Defining

$$W_v \stackrel{\rm def}{=} \{u \in C_v : |C_v \setminus \Gamma(u)| \geq 0.02 \cdot |\Gamma'(v)|\} \,,$$

[13] We note that the threshold for relative density is also different in the current case.

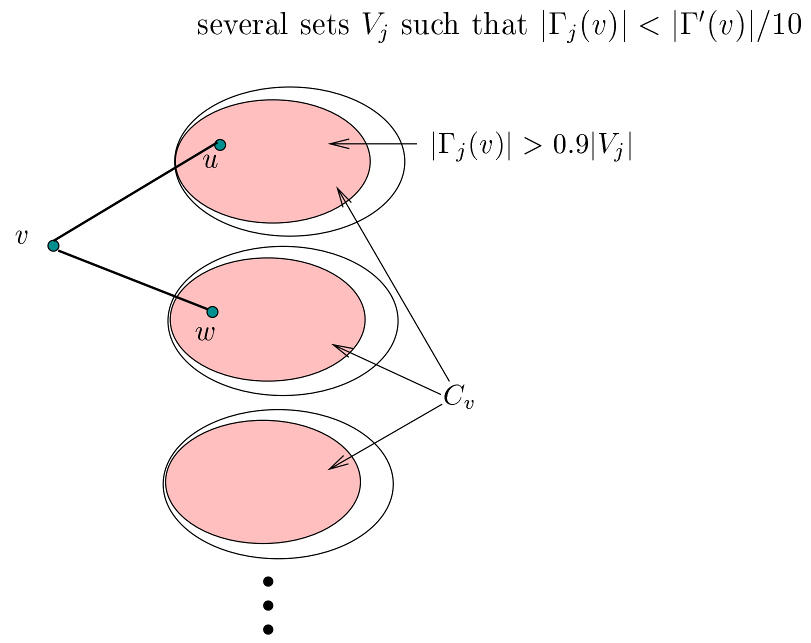

Figure 4: An Illustration for the proof of Claim 4.3.2, Case 2.1.

we note that $|W_v| \geq 0.02 \cdot |\Gamma'(v)|$. Next, we let $W_{v,u}$ be a $0.02 \cdot |W_v|$-size random subset of $C_v \setminus \Gamma(u) \subseteq \Gamma'(v) \setminus \Gamma(u)$. As in the previous cases, Part 1 follows by the definition of these sets. (However, unlike in the other cases, here we have $w \in \Gamma'(v)$ (and it also holds that $w \notin \Gamma(u)$).)

To establish Part 2, we first note that, for any fixed w, the expected size of $U_w^{(2)}$ is upper-bounded by

$$\begin{aligned} \sum_{v\in[N]:C_v\ni w} \sum_{u\in W_v} \frac{0.02\cdot|W_v|}{|C_v\setminus\Gamma(u)|} &\leq \sum_{v\in[N]:\Gamma'(v)\ni w} \sum_{u\in W_v} \frac{0.02\cdot|C_v|}{0.02\cdot|C_v|} \\ &= \sum_{v\in\Gamma'(w)} |W_v| \end{aligned} \tag{9}$$

where the inequality uses $|C_v \setminus \Gamma(u)| \geq 0.02 \cdot |\Gamma'(v)|$ and $W_v \subseteq C_v \subseteq \Gamma'(v)$. Analogously to the previous cases, it follows that if some w satisfies $|U_w^{(2)}| > \epsilon^{4/3}N^2$, then $\sum_{v\in\Gamma'(w)} |W_v| > \epsilon^{4/3}N^2/2$. This implies that either $|\Gamma'(w)| > \epsilon^{2/3}N/2$ or there exists $v \in \Gamma'(w)$ such that $|W_v| > \epsilon^{2/3}N$. Thus, Part 2 holds in Case 2.1.

Case 2.2: $\sum_{j\in J\setminus J'} |\Gamma_j(v)| \geq 0.5 \cdot |\Gamma'(v)|$. Let $J'' \stackrel{\rm def}{=} J \setminus J' = \{j : 1 \leq |\Gamma_j(v)| \leq 0.9|V_j|\}$, and note that for $j \in J''$ (as considered in this case) it may be that $|\Gamma_j(v)| \ll |V_j|$ and consequently for $j_1 \neq j_2 \in J''$ it may hold that $E(\Gamma_{j_1}(v), \Gamma_{j_2}(v)) \approx |\Gamma_{j_1}(v)| \cdot |\Gamma_{j_2}(v)|$. More generally, redefining $C_v \stackrel{\rm def}{=} \bigcup_{j\in J''} \Gamma_j(v)$, it may be that $|E(C_v, C_v)| \approx \binom{|C_v|}{2}$, and so the approach of Case 2.1 may not work in general (although it will work in the first subcase). Letting $J''' \stackrel{\rm def}{=} \{j \in J'' : |V_j| \leq |\Gamma'(v)|/10\}$, we consider two subcases:

1. If $\sum_{j\in J'''} |\Gamma_j(v)| \geq 0.4 \cdot |\Gamma'(v)|$ then we redefine $C_v \stackrel{\rm def}{=} \bigcup_{j\in J'''} \Gamma_j(v)$ and show that $|E(C_v, C_v)| \leq 0.99\binom{|C_v|}{2}$. Once the latter fact is established, we reach a situation

as in Case 2.1 and proceed exactly as in that case. To show that $|E(C_v, C_v)| \leq 0.99\binom{|C_v|}{2}$, we note that otherwise one obtains a contradiction to the optimality of the partition (by replacing the sub-partition $(V_j)_{j\in J'''}$ with $(C_v, (V_j \backslash C_v)_{j\in J'''})$, where $V_j \setminus C_v = \overline{\Gamma}_j(v)$). Details follow.
Assuming, towards the contradiction that $|E(C_v, C_v)| > 0.99\binom{|C_v|}{2}$, we lowerbound the gain from the aforementioned replacement as follows. The gain from edges inside C_v that do not connect vertices in the same V_j is lower-bounded by $0.99 \cdot \binom{|C_v|}{2} - \frac{|C_v|}{0.1|\Gamma'(v)|} \cdot \binom{0.1|\Gamma'(v)|}{2}$, which is lower-bounded by $0.36 \cdot |C_v|^2$ (when using $|\Gamma'(v)| \leq 2.5 \cdot |C_v|$). On the other hand, we upper-bound the loss from missing edges inside C_v and from superfluous edges introduced between C_v and the various sets V_j by $0.01 \cdot \binom{|C_v|}{2} + |C_v| \cdot \max_{j\in J'''}\{|V_j|\}$, which is upper-bounded by $0.26 \cdot |C_v|^2$ (when using $|V_j| \leq 0.1 \cdot |\Gamma'(v)| \leq 0.25 \cdot |C_v|$).

2. If $\sum_{j\in J''\setminus J'''} |\Gamma_j(v)| \geq 0.1 \cdot |\Gamma'(v)|$ then we proceed similarly to Case 1.1. Specifically, we define

$$W_v \stackrel{\text{def}}{=} \bigcup_{j\in J''\setminus J'''} \left\{ u \in \Gamma_j(v) : |\Gamma_j(u) \cap \overline{\Gamma}_j(v)| \geq \frac{|\overline{\Gamma}_j(v)|}{4} \right\}$$

and note that $W_v \subseteq \Gamma'(v)$ and that for every $j \in J'' \setminus J'''$ it holds that $|W_v \cap V_j| \geq |\Gamma_j(v)|/4$ (since $E(\Gamma_j(v), V_j \setminus \Gamma_j(v)) \geq |\Gamma_j(v)| \cdot |V_j \setminus \Gamma_j(v)|/2$). Using the subcase hypothesis, it follows that $|W_v| \geq \sum_{j\in J''\setminus J'''} |\Gamma_j(v)|/4 \geq |\Gamma'(v)|/40$, and using $j \in J'' \setminus J'''$ every $u \in W_v$ satisfies $|\Gamma_j(u) \cap \overline{\Gamma}_j(v)| \geq |\overline{\Gamma}_j(v)|/4 \geq |V_j|/40 \geq |\Gamma'(v)|/400$.
Next, for every $j \in J'' \setminus J'''$ and every $u \in W_v \cap V_j$, we define $W_{v,u}$ to be a random subset of size $|\Gamma'(v)|/400$ of $\Gamma_j(u) \cap \overline{\Gamma}_j(v)$. Indeed, for every $u \in W_v$ and $w \in W_{v,u}$ it holds that $w \notin \Gamma'(v)$ and $w \in \Gamma(u) \setminus \Gamma'(u)$. For an illustration, see Figure 5. Given the lower bounds on the sizes of the sets W_v and $W_{v,u}$, Part 1 follows.
To establish Part 2, we first note that, for any fixed $w \in V_j$, the expected size of $U_w^{(2)}$ is upper-bounded by

$$\begin{aligned}
\sum_{v\in[N]\setminus V_j} \sum_{u\in W_v\cap V_j} \frac{|\Gamma'(v)|/400}{|\Gamma_j(u) \cap \overline{\Gamma}_j(v)|} &\leq \sum_{v\in[N]\setminus V_j} \sum_{u\in\Gamma_j(v)} \frac{|\Gamma'(v)|}{10|V_j|} \\
&= \sum_{v\in[N]\setminus V_j} \frac{|\Gamma_j(v)| \cdot |\Gamma'(v)|}{10|V_j|}
\end{aligned}$$

where the inequality uses $|\Gamma_j(u) \setminus \Gamma_j(v)| \geq |V_j \setminus \Gamma_j(v)|/4 \geq |V_j|/40$. Analogously to the previous cases, it follows that if some $w \in V_j$ satisfies $|U_w^{(2)}| > \epsilon^{4/3}N^2$, then $\sum_{v\in[N]\setminus V_j} |\Gamma'(v)| \cdot |\Gamma_j(v)| > 5\epsilon^{4/3}N^2|V_j|$, which implies that either for some $v \in [N] \setminus V_j$ it holds that $|\Gamma'(v)| > \epsilon^{2/3}N$ or that $\sum_{v\in[N]\setminus V_j} |\Gamma_j(v)| > \epsilon^{2/3}N|V_j|$. In the latter case, there must be a vertex $u \in V_j$ such that $|\Gamma'(u)| > \epsilon^{2/3}N$. Thus, Part 2 holds in this subcase of Case 2.2.

Thus, we have established the claim for all subcases of Case 2.2.

Having completed the treatment of the two complementary cases of Case 2 (i.e., Cases 2.1 and 2.2), we complete the treatment of Case 2.

This completes the proof of Parts 1 and Part 2. Note that in each of the various cases we had $|W_{v,u}| \geq |W_v|/400$ (with the minimum lowerbound established in the second subcase of Case 2.2, where we used $|W_{v,u}| \geq |\Gamma'(v)|/400$).

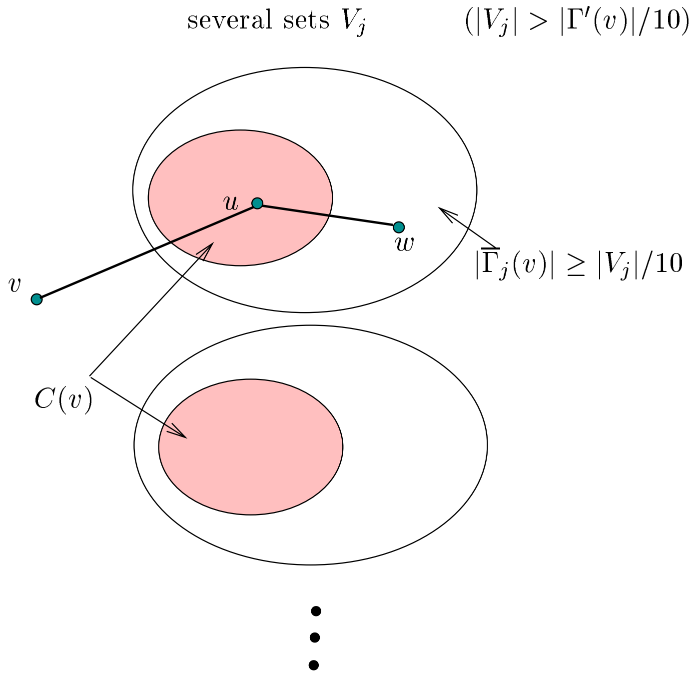

Figure 5: An Illustration for the proof of Claim 4.3.2, 2nd subcase of Case 2.2.

We now turn to proving Part 3. Except for Case 2.1, the modifications of the sets W_v and $W_{v,u}$ are analogous to those performed in the proof of Claim 4.3.1. Specifically, we first modify the sets W_v, by omitting from each W_v all vertices in $F(v)$ (recall that $F(v) = \{u : (v, u) \in F\}$). Note that we have decreased $\sum_v |W_v|$ by at most $2|F|$. The only case in which we make further modifications to the sets W_v is in Case 2.1. As we show subsequently, this causes a further decrease in $\sum_v |W_v|$ of at most $98|F|$. Hence, Eq. (7) follows by using the fact that Eq. (6) holds for the original sets W_v. Next, we modify the sets $W_{v,u}$, by omitting from each $W_{v,u}$ a few elements, selected at random, such that $|W_{v,u}| = |W_v|/400$ holds (for the modified sets). (This modification is done in order to allow the extension of the argument used in Part 2.)

To see that the generalized Part 2 holds too, we note that in all cases (including Case 2.1) the argument relies on the fact that $W_{v,u}$ is a random $\Omega(|W_v|)$-size subset of some (case-specific) subset of $\Gamma'(v)$ and on identifying a vertex v' for which $\Gamma'(v')$ is large (if some $U_w^{(2)}$ is large). The same applies to the modified sets (i.e., W_v's and $W_{v,u}$'s), however here we need to show that $\Gamma'(v') \setminus F(v')$ is large. Inspecting the various cases, we note that in all cases (except for Case 2.1) the original argument goes through. Specifically:

In Case 1.1 we showed that the existence of $w \in V_j$ such that $|U_w^{(2)}| > \epsilon^{4/3} N^2$ implies either the existence of $v \in V'$ (i.e., v satisfying $\xi(v) = j$) such that $|W_v| > \epsilon^{2/3} N/2$ or the existence of $u \in V_j$ such that $|\Gamma'(u)| > \epsilon^{2/3} N$. The same argument can be applied to the modified sets W_v and $W_{v,u}$, when replacing $E(V', V_j)$ by $E(V', V_j) \setminus F$ in Eq. (8). Thus, the first subcase implies that $|W_v| > \epsilon^{2/3} N/2$ (for some $v \in V'$ (and we are done since $|\Gamma'(v) \setminus F(v)| \geq |W_v|$)), whereas the second subcase implies the existence of $u \in V_j$ such that $|\Gamma'(u) \setminus F(u)| > \epsilon^{2/3} N$ (by using $|E(V', V_j) \setminus F| > |V_j| \cdot \epsilon^{2/3} N$, which implies the existence of $u \in V_j$ such that $|(\Gamma'(u) \setminus F(u)) \cap V'| > \epsilon^{2/3} N$).

In Case 1.2 we showed that the existence of $w \in V_i$ such that $|U_w^{(2)}| > \epsilon^{4/3}N^2$ implies the existence of $v \in V_i$ such that $|W_v| > \epsilon^{2/3}N$. The same argument applies to the modified sets W_v and $W_{v,u}$.

In Case 2.2 we reduced the first subcase to Case 2.1, whereas the second subcase was similar to Case 1.1. The adaptation is accordingly.

Indeed, this leaves us with Case 2.1, which is different from the other cases in the sense that it refers to sets $\Gamma'(w)$ such that the vertex w is not necessarily in some set W_v. Specifically, recall that in Case 2.1 we showed that the existence of $w \in V_j$ such that $|U_w^{(2)}| > \epsilon^{4/3}N^2$ implies that $\sum_{v\in\Gamma'(w)} |W_v| > \epsilon^{4/3}N^2/2$, which in turn implies that either $|\Gamma'(w)| > \epsilon^{2/3}N/2$ or $|W_v| > \epsilon^{2/3}N$ for some $v \in \Gamma'(w)$. However, unlike in Case 1.1,[14] we cannot replace $\Gamma'(w)$ by $\Gamma'(w) \setminus F(w)$, because $(v,u) \in U_w^{(2)}$ does not imply that $v \in \Gamma'(w) \setminus F(w)$. The source of trouble is that $W_{v,u}$ is selected with no reference to F.

The problem is resolved by modifying the selection of $W_{v,u}$ as follows. If $|F(v)| > |W_v|/98$ then W_v is reset to an empty set, and otherwise $W_{v,u}$ is selected as a random $(|W_v|/100)$-size subset of $(C_v \setminus \Gamma(u)) \setminus F(v) \subseteq \Gamma'(v) \setminus F(v)$ (rather than as a random $(|W_v|/50)$-size subset of $C_v \setminus \Gamma(u)$). This allows for replacing $\Gamma'(v) \ni w$ by $(\Gamma'(v) \setminus F(v)) \ni w$ in Eq. (9), and so we get

$$\begin{aligned}\sum_{v\in[N]:C_v\ni w}\ \sum_{u\in W_v} \frac{0.01\cdot|W_v|}{|(C_v\setminus\Gamma(u))\setminus F(v)|} &\leq \sum_{v\in[N]:(\Gamma'(v)\setminus F(v))\ni w}\ \sum_{u\in W_v} \frac{0.01\cdot|C_v|}{0.01\cdot|C_v|}\\ &= \sum_{v\in\Gamma'(w)\setminus F(w)} |W_v|\end{aligned}$$

where the inequality uses $|(C_v \setminus \Gamma(u)) \setminus F(v)| \geq 0.01 \cdot |\Gamma'(v)|$ and $W_v \subseteq C_v \subseteq \Gamma'(v) \setminus F(v)$. We conclude that the existence of $w \in V_j$ such that $|U_w^{(2)}| > \epsilon^{4/3}N^2$ implies that $\sum_{v\in\Gamma'(w)\setminus F(w)} |W_v| > \epsilon^{4/3}N^2/2$, which in turn implies that either $|\Gamma'(w) \setminus F(w)| > \epsilon^{2/3}N/2$ or $|W_v| > \epsilon^{2/3}N$ for some $v \in \Gamma'(w)\setminus F(w)$. Thus, Part 2 follows. We need, however, to examine the effect of this modification (of the sets $W_{v,u}$) on Part 1. The key observation is that the sum of the sizes of the W_v's decreases at most by $98|F|$, because the case of $|F(v)| > |W_v|/98$ (where W_v is reset to empty) causes a loss of at most $|W_v| < 98|F(v)|$, whereas the case of $|F(v)| \leq |W_v|/98$ (in which we avoid $F(v)$) causes (as usual) a loss of at most $|F(v)|$). This completes the treatment of general F, and the claim follows. $\square$

On the existence of effective witnesses. Combining the lemma's hypothesis with (the basic parts of) Claims 4.3.1 and 4.3.2, we infer the existence of $\Omega(\epsilon^2 N^3)$ witnesses. Moreover, the elaborate parts of these claims provide us with some structure that will be useful towards proving that (with high probability) the sample taken by Algorithm 4.2 contains at least one effective witness (i.e., a witness whose three vertex-pairs are inspected by the algorithm). Specifically, by the lemma's hypothesis, either $\sum_{v\in[N]} |\overline{\Gamma}(v)| \geq 0.001 \cdot \epsilon \cdot N^2$ or $\sum_{v\in[N]} |\Gamma'(v)| \geq 0.999 \cdot \epsilon \cdot N^2$. We first analyze the former case (i.e., $\sum_{v\in[N]} |\overline{\Gamma}(v)| \geq 0.001 \cdot \epsilon \cdot N^2$) and the treatment of the latter case (i.e., $\sum_{v\in[N]} |\Gamma'(v)| \geq 0.999 \cdot \epsilon \cdot N^2$) will follow (and be analogous). We consider two subcases:

[14]The crucial difference is that in Case 1.1 we considered $\Gamma'(u)$ for $(v,u) \in U_w^{(2)}$, which means that the modification of W_v allows replacing $\Gamma'(u)$ by $\Gamma'(u) \setminus F(u)$ (because $(v,u) \in U_w^{(2)}$ for the modified sets W_v implies that $v \in \Gamma'(u) \setminus F(u)$).

1. If $\sum_{v\in[N]:|\overline{\Gamma}(v)|\geq\epsilon^{2/3}N/2}|\overline{\Gamma}(v)| \geq 0.0001\cdot\epsilon\cdot N^2$ then applying Claim 4.3.1 with $F=\emptyset$ we obtain sets W_v's and $W_{v,u}$'s such that Part 1 of Claim 4.3.1 holds. In particular, it follows that

$$\begin{aligned}\sum_{v\in[N]:|W_v|\geq\epsilon^{2/3}N/8}|W_v| &\geq \sum_{v\in[N]:|\overline{\Gamma}(v)|\geq\epsilon^{2/3}N/2}\frac{|\overline{\Gamma}(v)|}{4}\\ &\geq \frac{0.0001\cdot\epsilon\cdot N^2}{4} = \Omega(\epsilon\cdot N^2).\end{aligned}$$

Recall that $\ell=\log_2(1/\epsilon)$. Thus, there exists $k\in\{1,...,(2\ell/3)+3\}$ such that for $V'\stackrel{\rm def}{=}\{v\in[N]:2^{-k}N\leq|W_v|<2^{-k+1}N\}$ it holds that $\sum_{v\in V'}|W_v|=\Omega(\epsilon\cdot N^2/\ell)$. Fixing this k, we note that $|V'|=\Omega(2^k\epsilon\cdot N/\ell)$ and thus $\Pr[R_k\cap V'\neq\emptyset]>8/9$, where R_k is as selected in Step 2 of Algorithm 4.2 (i.e., R_k is a random set of size $\Omega((2^k\epsilon/\ell)^{-1})$). Fixing any $v\in R_k\cap V'$, we have $|W_v|\geq 2^{-k}N$ and so $\Pr[S_k\cap W_v\neq\emptyset]>8/9$, where S_k is also as selected in Step 2 (i.e., S_k is a random set of size $\Omega(2^k)$). Finally, fixing any $u\in S_k\cap W_v$, we have $\Pr[S_k\cap W_{v,u}\neq\emptyset]>8/9$. Noting that all pairs $(R_k\times S_k)\cup(S_k\times S_k)$ are inspected by Algorithm 4.2, the claim follows.

2. If $\sum_{v\in[N]:|\overline{\Gamma}(v)|\geq\epsilon^{2/3}N/2}|\overline{\Gamma}(v)| < 0.0001\cdot\epsilon\cdot N^2$ then applying Claim 4.3.1 with $F=\{\{u,v\}: |\overline{\Gamma}(v)|\geq\epsilon^{2/3}N/2\}$ we obtain sets W_v's and $W_{v,u}$'s such that Claim 4.3.1 holds. In particular (by Part 1), it follows that

$$\begin{aligned}\sum_{v\in[N]}|W_v| &\geq \sum_{v\in[N]:|\overline{\Gamma}(v)|<\epsilon^{2/3}N/2}|W_v|\\ &\geq \sum_{v\in[N]:|\overline{\Gamma}(v)|\geq\epsilon^{2/3}N/2}\frac{|\overline{\Gamma}(v)|}{4}-2|F|\\ &\geq \left(\frac{0.001-0.0001}{4}-2\cdot 0.0001\right)\cdot\epsilon\cdot N^2 = \Omega(\epsilon\cdot N^2),\end{aligned}$$

whereas $|W_v|\leq|\overline{\Gamma}(v)\setminus F(v)|<\epsilon^{2/3}N/2$ holds for every $v\in[N]$. Note that we may assume, without loss of generality, that $|W_{v,u}|\leq|W_v|$ holds for every $u\in W_v$. (Actually, $|W_{v,u}|=|W_v|/4$ holds for the sets constructed in the proof of Claim 4.3.1.)

Letting $U_w^{(1)}\stackrel{\rm def}{=}\{v: w\in W_v\}$, for every w it holds that $|U_w^{(1)}|<\epsilon^{2/3}N/2$ (because $v\in U_w^{(1)}$ implies $w\in\overline{\Gamma}(v)$ and $(v,w)\notin F$). Also, by Part 2, we get $|U_w^{(2)}|<\epsilon^{4/3}N$ for every w. Using the following Claim 4.3.3, we shall show that in such a case (with high probability) the sample S selected in Step 1 (of Algorithm 4.2) contains a witness (i.e., a triple (v,u,w) such that $u\in W_v$ and $w\in W_{v,u}$). Loosely speaking, the expected number of witnesses exceeds any constant, whereas the upper-bounds on the sets $|W_v|$, $|U_v^{(1)}|$ and $|U_v^{(2)}|$ guarantees sufficient concentration around the expected value.

The treatment of the case in which $\sum_{v\in[N]}|\Gamma'(v)|\geq 0.999\cdot\epsilon\cdot N^2$ is analogous. Specifically, we consider analogous subcases (with different constants in the differentiating thresholds) and invoke Claim 4.3.2. Either way, the analysis of the second subcase (above) relies on the following claim.

Claim 4.3.3 (sampling triples via a 3-way Cartesian product of samples): *Suppose that the following conditions hold:*

1. $\sum_{v\in[N]}\sum_{u\in W_v}|W_{v,u}|=\Omega(\epsilon^2\cdot N^3)$

2. For every $v \in [N]$, it holds that $\max(|W_v|, |U_v^{(1)}|, |U_v^{(2)}|) < \epsilon^{2/3}N$, *where* $U_v^{(1)} \stackrel{\rm def}{=} \{x : v \in W_x\}$ *and* $U_v^{(2)} \stackrel{\rm def}{=} \{(x,y) : v \in W_{x,y}\}$.

3. For every $v \in [N]$ and $u \in W_v$, it holds that $|W_{v,u}| < \epsilon^{2/3}N$.

Then, for a sufficiently large constant c that depends only on the constant in the O-notation, with probability at least $2/3$, a uniformly selected sample of $c \cdot \epsilon^{-2/3}$ vertices contains a triple (v,u,w) such that $u \in W_v$ and $w \in W_{v,u}$.

Recall that we only invoke Claim 4.3.3 in the second forgoing case, and whenever we do so all the conditions in the hypothesis hold. Specifically, we have $\sum_{v\in[N]}\sum_{u\in W_v}|W_{v,u}| = \sum_{v\in[N]}\Omega(|W_v|^2) = \Omega(\epsilon^2 \cdot N^3)$ (since $\sum_{v\in[N]}|W_v| = \Omega(\epsilon \cdot N^2)$) as well as $|W_v|, |W_{v,u}|, |U_v^{(1)}| < \epsilon^{-2/3}N$ (since $W_v \subseteq \overline{\Gamma}(v) \setminus F(v)$ (or $W_v \subseteq \Gamma'(v) \setminus F(v)$) and the same holds for $U_v^{(1)}$). Furthermore, Claim 4.3.1 (resp., Claim 4.3.2) implies that in this case (where $|\overline{\Gamma}(v) \setminus F(v)| < \epsilon^{-2/3}N/2$ (resp., $|\Gamma'(v) \setminus F(v)| < \epsilon^{-2/3}N/2$), it holds that $|U_v^{(2)}| \le 10\epsilon^{-4/3}N^2$. By replacing ϵ with $\epsilon/10$, the hypothesis holds.

Proof: We may assume, without loss of generality, that for any v and $u \in W_v$ it holds that $|W_{v,u}| \le |W_v|$. (Note that this is the case anyhow in the proofs of Claims 4.3.1 and 4.3.2.) We denote the vertices of the sample S by $v_1, \ldots, v_s, u_1, \ldots, u_s, w_1, \ldots, w_s$. We shall prove that, with probability at least $1 - O(s^{-1}\epsilon^{-2/3})$, there exists a triple $(i,j,k) \in [s]^3$ such that $u_j \in W_{v_i}$ and $w_k \in W_{v_i,w_j}$. The proof boils down to applying Chebyshev's Inequality to $\sum_{i,j,k\in[s]}\zeta_{i,j,k}$, where $\zeta_{i,j,k} = 1$ if $u_j \in W_{v_i}$ and $w_k \in W_{v_i,u_j}$, and $\zeta_{i,j,k} = 0$ otherwise. We first note that

$$\begin{aligned}
\mu &\stackrel{\rm def}{=} \mathrm{Exp}_S\left[\sum_{i,j,k\in[s]}\zeta_{i,j,k}\right] \\
&= s^3 \cdot \mathrm{Pr}_{v,u,w\in[N]}[u \in W_v \ \wedge \ w \in W_{v,u}] \\
&= s^3 \cdot \frac{1}{N^3} \cdot \sum_{v\in[N]}\sum_{u\in W_v}|W_{v,u}| \\
&= \Omega(s^3 \cdot \epsilon^2)
\end{aligned}$$

where the last line follows by the first condition in the hypothesis. By Chebyshev's Inequality it follows that

$$\begin{aligned}
\Pr\left[\sum_{i,j,k\in[s]}\zeta_{i,j,k} = 0\right] &\le \frac{\mathrm{Var}[\sum_{i,j,k\in[s]}\zeta_{i,j,k}]}{\mathrm{Exp}[\sum_{i,j,k\in[s]}\zeta_{i,j,k}]^2} \\
&= \mu^{-2} \cdot \left(\mathrm{Exp}\left[\left(\sum_{i,j,k\in[s]}\zeta_{i,j,k}\right)^2\right] - \mathrm{Exp}\left[\sum_{i,j,k\in[s]}\zeta_{i,j,k}\right]^2\right) \\
&= \mu^{-2} \cdot \left(\left(\sum_{\overline{\ell}\in[s]^6}\mathrm{Exp}[\zeta_{i_1,j_1,k_1} \cdot \zeta_{i_2,j_2,k_2}]\right) - \mu^2\right) \qquad (10)
\end{aligned}$$

where $\overline{\ell} = (i_1, i_2, j_1, j_2, k_1, k_2)$. The upper bounds on $|W_v|, |W_{v,u}|, |U_w^{(1)}|$ and $|U_w^{(2)}|$ will be used in upper-bounding the large sum (i.e., $\sum_{\overline{\ell}\in[s]^6}\mathrm{Exp}[\zeta_{i_1,j_1,k_1} \cdot \zeta_{i_2,j_2,k_2}]$). We decompose the latter sum into partial sums that correspond to the following cases (regarding the relations between i_1-vs-i_2, j_1-vs-j_2, and k_1-vs-k_2).

Case of $i \stackrel{\rm def}{=} i_1 = i_2$*,* $j \stackrel{\rm def}{=} j_1 = j_2$*, and* $k \stackrel{\rm def}{=} k_1 = k_2$*.* There are s^3 such terms, each having value $\mathrm{Exp}[\zeta_{i,j,k}^2] = \mathrm{Exp}[\zeta_{i,j,k}]$, which equals $\Pr_{v,u,w \in [N]}[u \in W_v \wedge w \in W_{v,u}] = \mu/s^3$. Thus, the total contribution of this case is μ.

Case of $i \stackrel{\rm def}{=} i_1 = i_2$*,* $j \stackrel{\rm def}{=} j_1 = j_2$*, and* $k_1 \neq k_2$*.* There are less than s^4 such terms, each having value $\mathrm{Exp}[\zeta_{i,j,k_1} \cdot \zeta_{i,j,k_2}]$, which equals

$$\begin{aligned}
&\Pr_{v,u,w_1,w_2 \in [N]}[u \in W_v \ \wedge \ w_1, w_2 \in W_{v,u}] \\
&\quad \leq \ \Pr_{v,u,w_1 \in [N]}[u \in W_v \ \wedge \ w_1 \in W_{v,u}] \cdot \max_{v,u,w_1 \in [N]} \left\{ \Pr_{w_2 \in [N]}[w_2 \in W_{v,u}] \right\} \\
&\quad < \ \frac{\mu}{s^3} \cdot \epsilon^{2/3}
\end{aligned}$$

where the inequality is due to $|W_{v,u}| < \epsilon^{2/3} N$. Thus, the total contribution of this case is smaller than $(s\epsilon^{2/3}) \cdot \mu$.

Case of $i \stackrel{\rm def}{=} i_1 = i_2$*,* $j_1 \neq j_2$*, and* $k \stackrel{\rm def}{=} k_1 = k_2$*.* There are less than s^4 such terms, each having value $\mathrm{Exp}[\zeta_{i,j_1,k} \cdot \zeta_{i,j_2,k}]$, which equals

$$\begin{aligned}
&\Pr_{v,u_1,u_2,w \in [N]}[u_1, u_2 \in W_v \ \wedge \ w \in W_{v,u_1} \cap W_{v,u_2}] \\
&\quad \leq \ \Pr_{v,u_1,w \in [N]}[u_1 \in W_v \ \wedge \ w \in W_{v,u_1}] \cdot \max_{v,u_1,w \in [N]} \left\{ \Pr_{u_2 \in [N]}[u_2 \in W_v] \right\} \\
&\quad < \ \frac{\mu}{s^3} \cdot \epsilon^{2/3}
\end{aligned}$$

where the inequality is due to $|W_v| < \epsilon^{2/3} N$. Thus, the total contribution of this case is smaller than $(s\epsilon^{2/3}) \cdot \mu$.

Case of $i \stackrel{\rm def}{=} i_1 = i_2$*,* $j_1 \neq j_2$*, and* $k_1 \neq k_2$*.* There are less than s^5 such terms, each having value $\mathrm{Exp}[\zeta_{i,j_1,k_1} \cdot \zeta_{i,j_2,k_2}]$, which equals

$$\begin{aligned}
&\Pr_{v,u_1,u_2,w_1,w_2 \in [N]}[u_1, u_2 \in W_v \ \wedge \ w_1 \in W_{v,u_1} \ \wedge \ w_2 \in W_{v,u_2}] \\
&\quad \leq \ \Pr_{v,u_1,w_1 \in [N]}[u_1 \in W_v \ \wedge \ w_1 \in W_{v,u_1}] \cdot \max_{v,u_1,w_1 \in [N]} \left\{ \Pr_{u_2,w_2 \in [N]}[u_2 \in W_v \ \wedge \ w_2 \in W_{v,u_2}] \right\} \\
&\quad < \ \frac{\mu}{s^3} \cdot (\epsilon^{2/3})^2
\end{aligned}$$

where the inequality is due to $|W_v| < \epsilon^{2/3} N$ and $|W_{v,u_2}| < \epsilon^{2/3} N$. Thus, the total contribution of this case is smaller than $(s\epsilon^{2/3})^2 \cdot \mu$.

Case of $i_1 \neq i_2$*,* $j \stackrel{\rm def}{=} j_1 = j_2$*, and* $k \stackrel{\rm def}{=} k_1 = k_2$*.* There are less than s^4 such terms, each having value $\mathrm{Exp}[\zeta_{i_1,j,k} \cdot \zeta_{i_2,j,k}]$, which equals

$$\begin{aligned}
&\Pr_{v_1,v_2,u,w \in [N]}[u \in W_{v_1} \cap W_{v_2} \ \wedge \ w \in W_{v_1,u} \cap W_{v_2,u}] \\
&\quad \leq \ \Pr_{v_1,u,w \in [N]}[u \in W_{v_1} \ \wedge \ w \in W_{v_1,u}] \cdot \max_{v_1,u,w \in [N]} \left\{ \Pr_{v_2 \in [N]}[u \in W_{v_2}] \right\} \\
&\quad < \ \frac{\mu}{s^3} \cdot \epsilon^{2/3}
\end{aligned}$$

where the inequality is due to $|U_u^{(1)}| < \epsilon^{2/3} N$ (and $u \in W_{v_2}$ iff $v_2 \in U_u^{(1)}$). Thus, the total contribution of this case is smaller than $(s\epsilon^{2/3}) \cdot \mu$.

Case of $i_1 \neq i_2$, $j_1 \neq j_2$, and $k \stackrel{\rm def}{=} k_1 = k_2$. There are less than s^5 such terms, each having value $\mathrm{Exp}[\zeta_{i_1,j_1,k} \cdot \zeta_{i_2,j_2,k}]$, which equals

$$\begin{aligned} & \Pr_{v_1,v_2,u_1,u_2,w \in [N]}[u_1 \in W_{v_1} \ \wedge \ u_2 \in W_{v_2} \ \wedge \ w \in W_{v_1,u_1} \cap W_{v_2,u_2}] \\ \leq \ & \Pr_{v_1,u_1,w \in [N]}[u_1 \in W_{v_1} \ \wedge \ w \in W_{v_1,u_1}] \cdot \max_{v_1,u_1,w \in [N]} \left\{ \Pr_{u_2,v_2 \in [N]}[w \in W_{v_2,u_2}] \right\} \\ < \ & \frac{\mu}{s^3} \cdot \epsilon^{4/3} \end{aligned}$$

where the inequality is due to $|U_w^{(2)}| < \epsilon^{4/3} N^2$ (and $w \in W_{v_2,u_2}$ iff $(v_2,u_2) \in U_w^{(2)}$). Thus, the total contribution of this case is smaller than $s^2 \epsilon^{4/3} \cdot \mu$.

Case of $i_1 \neq i_2$, $j \stackrel{\rm def}{=} j_1 = j_2$, and $k_1 \neq k_2$. There are less than s^5 such terms, each having value $\mathrm{Exp}[\zeta_{i_1,j,k_1} \cdot \zeta_{i_2,j,k_2}]$, which equals

$$\begin{aligned} & \Pr_{v_1,v_2,u,w_1,w_2 \in [N]}[u \in W_{v_1} \cap W_{v_2} \ \wedge \ w_1, w_2 \in W_{v_1,u} \cap W_{v_2,u}] \\ \leq \ & \Pr_{v_1,u,w_1 \in [N]}[u \in W_{v_1} \ \wedge \ w_1 \in W_{v_1,u}] \cdot \max_{v_1,u,w_1 \in [N]} \left\{ \Pr_{v_2,w_2 \in [N]}[u \in W_{v_2} \ \wedge \ w_2 \in W_{v_2,u}] \right\} \\ < \ & \frac{\mu}{s^3} \cdot \epsilon^{2/3} \end{aligned}$$

where the inequality is due to $|U_u^{(1)}| < \epsilon^{2/3} N$ and $|W_{v_2,u}| < \epsilon^{2/3} N$. Thus, the total contribution of this case is smaller than $(s\epsilon^{2/3})^2 \cdot \mu$.

Case of $i_1 \neq i_2$, $j_1 \neq j_2$, and $k_1 \neq k_2$. There are less than s^6 such terms, each having value $\mathrm{Exp}[\zeta_{i_1,j_1,k_1} \cdot \zeta_{i_2,j_j,k_2}] = \mathrm{Exp}[\zeta_{i,j,k}]^2$, which equals $(\mu/s^3)^2$. Thus, the total contribution of this case is smaller than μ^2.

Thus, we have one case (i.e., the first one) contributing μ, three cases (each) contributing $s\epsilon^{2/3} \cdot \mu$, three cases (each) contributing $(s\epsilon^{2/3})^2 \cdot \mu$, and one case (i.e., the last one) contributing μ^2. Using these upper bounds in Eq. (10), we obtain

$$\begin{aligned} \Pr\left[\sum_{i,j,k \in [s]} \zeta_{i,j,k} = 0\right] \ & < \ \mu^{-2} \cdot \left(\left(\mu + 3 \cdot s\epsilon^{2/3} \cdot \mu + 3 \cdot (s\epsilon^{2/3})^2 \cdot \mu + \mu^2 \right) - \mu^2 \right) \\ & = \ \mu^{-1} \cdot \left(1 + 3s\epsilon^{2/3} + 3(s\epsilon^{2/3})^2 \right). \end{aligned}$$

Using $\mu = \Omega(s^3 \epsilon^2)$ and a sufficiently large $s = O(\epsilon^{-2/3})$, we obtain an error bound of $O((s\epsilon^{2/3})^2/(s^3\epsilon^2)) = O(s^{-1}\epsilon^{-2/3}) < 1/3$, and the claim follows. □

This completes the proof of Lemma 4.3. ■

5 Larger Adaptive vs Non-adaptive Complexity Gaps

We start by establishing Theorem 1.2, which refers to the adaptive vs non-adaptive complexity gap of testing Bi-Clique Collections. We believe that the ideas underlying the adaptive algorithm and the non-adaptive lower-bound (presented in Sections 5.1 and 5.2) can serve as a basis for establishing the larger gap stated in Conjecture 1.3. Indeed, as shown in Section 5.3, this is the case with respect to the non-adaptive lower-bound (which indeed establishes Part 2 of Conjecture 1.3). In Section 5.4 we outline an adaptive algorithm that we believe to be suitable for Part 1 of Conjecture 1.3.

5.1 The Adaptive Query Complexity of Bi-Clique Collection

The tester for $\mathcal{BCC}$ is obtained by extending the ideas that underly the tester for $\mathcal{CC}$ (i.e., Algorithm 3.1). The extension is relatively straightforward, but the analysis will have to address additional difficulties (i.e., beyond those encountered in the analysis of Algorithm 3.1).

Algorithm 5.1 (adaptive tester for $\mathcal{BCC}$): *On input N and ϵ and oracle access to a graph $G = ([N], E)$, the tester sets $\ell = \log_2(1/\epsilon)+2$, $t_1 = \Theta(\ell)$ and $t_2 = \Theta(\ell^4)$, and proceeds in ℓ iterations as follows: For $i = 1, ..., \ell$, the tester selects uniformly $t_1 \cdot 2^i$ start vertices and for each selected vertex $v \in [N]$ performs the following sub-test, denoted* sub-test$_i(v)$*:*

1. *The sub-text selects at random a sample, S, of $t_2/(2^i\epsilon)$ vertices, and determines $\Gamma_S(v) = S \cap \Gamma(v)$, by making the queries (v, w) for each $w \in S$. If $\Gamma_S(v) \neq \emptyset$ then it selects u at random in $\Gamma_S(v)$ and continue to the following steps.* (Otherwise, the sub-test halts and accepts v.)
2. *The sub-text determines $\Gamma_S(u) = S \cap \Gamma(u)$, by making the queries (u, w) for each $w \in S$.*
3. *If $|\Gamma_S(v) \times \Gamma_S(u)| \leq t_2/2^i\epsilon$ then the sub-test checks that for every $(w_1, w_2) \in \Gamma_S(v) \times \Gamma_S(u)$ it holds that $(w_1, w_2) \in E$. Otherwise* (i.e., $|\Gamma_S(v) \times \Gamma_S(u)| > t_2/2^i\epsilon$), *it selects a sample of $t_2/(2^i\epsilon)$ pairs in $\Gamma_S(v) \times \Gamma_S(u)$ and checks that each selected pair is in E.*
4. *Let $B = (\Gamma_S(v) \times \Gamma_S(v)) \cup (\Gamma_S(u) \times \Gamma_S(u))$. If $|B| \leq t_2/2^i\epsilon$ then the sub-test checks that for every $(w_1, w_2) \in B$ it holds that $(w_1, w_2) \notin E$. Otherwise* (i.e., $|B| > t_2/2^i\epsilon$), *it selects a sample of $t_2/(2^i\epsilon)$ pairs in B and checks that each selected pair is in* not *E.*
5. *The sub-text selects a sample of $t_2/(2^i\epsilon)$ pairs in $(\Gamma_S(v) \cup \Gamma_S(u)) \times (S \setminus (\Gamma_S(v) \cup \Gamma_S(u)))$ and check that each selected pair is* not *in E.*

The sub-test (i.e., sub-test$_i(v)$) *accepts if and only if all checks were positive* (i.e., no edges were missed in Step 3 and no edges were detected in Steps 4 and 5). *The tester itself accepts if and only if all $\sum_{i=1}^{\ell} t_1 \cdot 2^i$ invocations of the sub-test accepted.*

The query complexity of this algorithm is $\sum_{i=1}^{\ell}(t_1 \cdot 2^i) \cdot O(t_2/2^i\epsilon) = O(\ell \cdot t_1 t_2/\epsilon) = \widetilde{O}(1/\epsilon)$. Clearly, this algorithm accepts (with probability 1) any graph that is in $\mathcal{BCC}$. It remains to analyze its behavior on graphs that are ϵ-far from $\mathcal{BCC}$.

Lemma 5.2 *If $G = ([N], E)$ is ϵ-far from $\mathcal{BCC}$, then on input N, ϵ and oracle access to G, Algorithm 5.1 rejects with probability at least $2/3$.*

Part 1 of Theorem 1.2 follows.

Proof: We proceed as in the proof of Lemma 3.2; that is, we will show that if Algorithm 5.1 accepts with probability at least $1/3$ then the graph is ϵ-close to $\mathcal{BCC}$. The proof evolves around a revised notion of i-good start vertices, which is defined on top of the notion of i-good edges. The definition refers to the parameters γ_2 and γ_3, which will be determined such that $\gamma_2 = \Theta(1/t_2)$ and $\gamma_1 \cdot \gamma_3 = \Theta(1/t_1)$.

Definition 5.2.1 *An edge (v, u) is i-*good *if the following three conditions hold.*

1. *The number of missing edges in $\Gamma(v) \times \Gamma(u)$ is at most $\gamma_2 \cdot 2^i \epsilon \cdot |\Gamma(v,u)| \cdot N$ edges, where $\Gamma(v,u) \stackrel{\rm def}{=} \Gamma(v) \cup \Gamma(u)$; that is, $|(\Gamma(v) \times \Gamma(u)) \setminus E| \leq \gamma_2 \cdot 2^i \epsilon \cdot |\Gamma(v,u)| \cdot N$.*

2. *The number of edges in $(\Gamma(v) \times \Gamma(v)) \cup (\Gamma(u) \times \Gamma(u))$ is at most $\gamma_2 \cdot 2^i \epsilon \cdot |\Gamma(v,u)| \cdot N$.*

3. *For every positive integer $j \leq j_0 \stackrel{\rm def}{=} \log_2(|\Gamma(v,u)|/(\gamma_2 \cdot 2^i \epsilon N))$, the number of vertices in $\Gamma(v,u)$ that have at least $\gamma_2 \cdot 2^{i+j} \epsilon \cdot N$ edges going out of $\Gamma(v,u)$ is at most $2^{-j} \cdot |\Gamma(v,u)|$.*

A vertex v is i-good if at least $(1-\gamma_3) \cdot |\Gamma(v)|$ of its neighbors yield a edge that is i-good; that is, if $|\{u \in \Gamma(v) : (v,u) \text{ is } i\text{-good}\}| \geq (1-\gamma_3) \cdot |\Gamma(v)|$.

Claim 5.2.2 *If v has degree at least $\gamma_2 \cdot 2^i \epsilon \cdot N$ and is not i-good, then the probability that* sub-test$_i(v)$ *rejects is at least $\gamma_3/2$.*

Proof: By the hypothesis $|\Gamma(v)| \geq \gamma_2 \cdot 2^i \epsilon \cdot N$, with probability at least 0.9, Step 1 of sub-test$_i(v)$ generates a non-empty sample of vertices in $\Gamma(v)$. Conditioned on this event (and using the hypothesis that v is not i-good), with probability at least γ_3, the vertex $u \in \Gamma(v)$ selected in this sample is such that (v,u) is not i-good. We fix such an edge (v,u) for the rest of this proof.

Assume that Condition 1 of i-goodness does not hold for (v,u), and let $\rho \stackrel{\rm def}{=} \frac{\gamma_2 \cdot 2^i \epsilon \cdot |\Gamma(v,u)| \cdot N}{|\Gamma(v)| \cdot |\Gamma(u)|} \geq \frac{\gamma_2 \cdot 2^i \epsilon \cdot N}{\min(|\Gamma(v)|, |\Gamma(u)|)}$ denote (the lower bound on) the fraction of missing edges in $\Gamma(v) \times \Gamma(u)$. (Note that this event may happen only if $\min(|\Gamma(v)|, |\Gamma(u)|) \geq \gamma_2 \cdot 2^i \epsilon \cdot N$.) Then, with probability at least 0.9, it holds that $\min(|\Gamma_S(v)|, |\Gamma_S(u)|) > m/2$, where $m \stackrel{\rm def}{=} \frac{t_2}{\epsilon 2^i} \cdot \frac{\min(|\Gamma(v)|, |\Gamma(u)|)}{N} \geq t_2 \cdot \gamma_2 \gg 1$. Also note that the members of $\Gamma_S(v)$ and $\Gamma_S(u)$ are distributed uniformly in $\Gamma(v)$ and $\Gamma(u)$, respectively. Considering $n = m/2$ uniformly distributed vertices in $\Gamma(v)$ and n uniformly distributed vertices in $\Gamma(u)$, it follows (as in the proof of Claim 3.2.2) that, with probability at least 0.9, the fraction of edges that are missing in the subgraph induced by the said sample is at least $\rho/2$. It follows that Step 3 rejects with probability at least $0.9^2 > 0.8$ (regardless if it examines all pairs in $\Gamma_S(v) \times \Gamma_S(u)$ or just examines a random sample of $\frac{t_2}{2^i \epsilon} \geq \frac{t_2 \gamma_2}{\rho}$ pairs).

The treatment of Condition 2 is similar, except that here we refer to the number of edges (in $(\Gamma(v) \times \Gamma(v)) \cup (\Gamma(u) \times \Gamma(u))$) over $|\Gamma(v)|^2 + |\Gamma(u)|^2 = \Theta(|\Gamma(v,u)|^2)$. Indeed, treating $\Gamma(v,u)$ as a whole facilitates the streamlining of the proof with the treatment of Condition 1 in Claim 3.2.2. We conclude that if Condition 2 (of i-goodness of (v,u)) is violated, then Step 4 of the test rejects with probability at least 0.8.

Finally, we turn to Condition 3 of i-goodness. Assuming that this condition does not hold for (v,u), we show that Step 5 of the test rejects with probability at least 0.8. The proof is analogous to the analysis of Condition 2 in Claim 3.2.2, except that $\Gamma(v,u)$ replaces $\Gamma(v)$. Thus, sub-test$_i(v)$ rejects with probability at least $0.9 \cdot \gamma_2 \cdot 0.8$, and the current claim follows. □

Claim 5.2.3 *If Algorithm 5.1 accepts with probability at least $1/3$ then for every $i \in [\ell]$ the number of vertices of degree at least $\gamma_2 \cdot 2^i \epsilon \cdot N$ that are not i-good is at most $\gamma_1 \cdot 2^{-i} \cdot N$, where $\gamma_1 \gamma_3 = \Theta(1/t_1)$.*

Proof: Assuming to the contrary that the number of these vertices exceeds $\gamma_1 \cdot 2^{-i} \cdot N$, Claim 5.2.2 implies that a single invocation of sub-test$_i$ rejects with probability at least $\gamma_1 2^{-i} \cdot \gamma_3/2$. Recalling that Algorithm 5.1 invokes sub-test$_i$ on $t_1 \cdot 2^i$ random vertices (and using $t_1 \geq 2 \cdot (\gamma_1 \gamma_3)^{-1}$), the claim follows. □

Additional difficulties. As stated up-front, the current proof faces additional difficulties that were not encountered in the proof of Lemma 3.2. These difficulties refer to the partition reconstruction procedure, which is supposed to provide an approximately good partition of the graph to bi-cliques. The first problem refers to the case that (v,u) is i-good, but most of $\Gamma(v,u)$ belongs to previously identified bi-cliques and furthermore these vertices reside in $\Gamma(u)$ (rather than in $\Gamma(v)$). Thus, we cannot "charge" these vertices to edges that are adjacent to v, but rather develop a charging rule that allows us to charge v indirectly via its typical neighbors u. The second problem refers to the treatment of low-degree vertices, and it arises from the fact that vertices in $\Gamma(v,u)$ may have vastly different degrees (which, indeed, occurs in the case that $\Gamma(v)$ has a significantly different cardinality than $\Gamma(u)$). Our solution is based on using two different degree thresholds (depending on the relation between the degree of a vertex and the degree of most of its neighbors). With this motivation in mind, we turn to the actual description of the (iterative) partition-reconstruction procedure.

The partition reconstruction procedure. The iterative procedure is initiated with $C = L_0 = L_0^{(1)} = L_0^{(2)} = L_0^{(I)} = \emptyset$, $R_0 = [N]$ and $i = 1$, where C denotes the set of vertices "covered" (by bi-cliques) so far, R_{i-1} denotes the set of "remaining" vertices after iteration $i-1$ and L_{i-1} denotes the set of vertices cast aside (as having "low degree") in iteration $i-1$. The set L_{i-1} is the union of three sets, $L_{i-1}^{(1)}$, $L_{i-1}^{(2)}$, and $L_{i-1}^{(I)}$, where the first two sets correspond to two degree thresholds, denoted β_1 and β_2, and the third set consists of many subsets that use intermediate thresholds (for avoiding a non-smooth transition). (We shall set $\beta_1 = \Theta(1/\ell)$ and $\beta_2 = \Theta(\beta_1/\ell) \gg \gamma_2$.) The $i^{\rm th}$ iteration proceeds as follows, where $i = 1, ..., \ell$ and F_i is initialized to $\emptyset$.

1. Pick an arbitrary vertex $v \in R_{i-1} \setminus C$ that satisfies the following three conditions
   (a) v is i-good.
   (b) v has sufficiently high degree in the following sense: either $|\Gamma(v)| \geq \beta_1 \cdot 2^i \epsilon \cdot N$ or for some $k \in [\ell']$, where $\ell' = \log_{0.9}(\beta_2/\beta_1) = O(\log \ell)$, both $|\Gamma(v)| \geq 0.9^k \cdot \beta_1 \cdot 2^i \epsilon \cdot N$ and $\phi_k(v)$ hold, where $\phi_k(v)$ represents the condition that a significant fraction of v's neighbors have a significantly higher degree than v itself; specifically, $\phi_k(v)$ holds if
   $$\left|\left\{w \in \Gamma(v) : |\Gamma(w)| > \left(1.1 + \frac{k}{10\ell'}\right) \cdot |\Gamma(v)|\right\}\right| \;>\; \frac{|\Gamma(v)|}{100\ell\ .} \tag{11}$$
   Note that $\phi_{\ell'}(v)$ holds if $|\{w \in \Gamma(v) : |\Gamma(w)| > 1.2 \cdot |\Gamma(v)|\}|$ is greater than $|\Gamma(v)|/100\ell$, and the corresponding degree bound is $\beta_2 \cdot 2^i \epsilon \cdot N$ (because $0.9^{\ell'} = \beta_2/\beta_1$).
   (c) There exists $u \in \Gamma(v) \setminus C$ such that the edge (v,u) is i-good and
   $$\left|(\Gamma(v,u) \setminus C) \setminus \left(\bigcup_{j \leq i-1} L_j\right)\right| \;\geq\; \frac{|\Gamma(v,u)|}{5}$$
   (i.e., relatively few vertices of $\Gamma(v,u)$ are covered by C or cast aside in previous iterations due to having low degree).

If no such vertex v exists, then define

$$\begin{aligned}
L_i^{(1)} &= \{v \in R_{i-1} \setminus C : \neg\phi_1(v) \wedge (|\Gamma(v)| < \beta_1 \cdot 2^i \epsilon \cdot N)\}, \\
L_i^{(I)} &= \bigcup_{k \in [\ell'-1]} \{v \in R_{i-1} \setminus C : \phi_k(v) \wedge \neg\phi_{k+1}(v) \wedge (|\Gamma(v)| < 0.9^k \beta_1 \cdot 2^i \epsilon \cdot N)\}, \\
L_i^{(2)} &= \{v \in R_{i-1} \setminus C : \phi_{\ell'}(v) \wedge (|\Gamma(v)| < \beta_2 \cdot 2^i \epsilon \cdot N)\},
\end{aligned}$$

$L_i = L_i^{(1)} \cup L_i^{(I)} \cup L_i^{(2)}$, and $R_i = R_{i-1} \setminus (L_i \cup C)$.

If $i < \ell$ then proceed to the next iteration, and otherwise terminate.

2. For vertex v as selected in Step 1, pick an arbitrary $u \in \Gamma(v) \setminus C$ satisfying Condition 1c. Let $C_{v,u} = \{w \in \Gamma(v,u) : |\Gamma(w) \setminus \Gamma(v,u)| < |\Gamma(v,u)|\}$. Form a new bi-clique with the vertex set $C'_{v,u} \leftarrow C_{v,u} \setminus C$, and update $F_i \leftarrow F_i \cup \{(v,u)\}$ and $C \leftarrow C \cup C'_{v,u}$. This bi-clique will have $\Gamma'(v) \stackrel{\text{def}}{=} \Gamma(v) \cap C'_{v,u}$ on one side and $\Gamma'(u) \stackrel{\text{def}}{=} \Gamma(u) \cap C'_{v,u}$ on the other side.

Note that by Condition 1c (and the definition of i-goodness), for every $(v,u) \in F_i$, it holds that $|C_{v,u}| > (1-o(1)) \cdot |\Gamma(v,u)|$ and $|\Gamma(v,u) \setminus C| \geq |\Gamma(v,u)|/5$. Thus, $|C'_{v,u}| \geq |C_{v,u}| - |\Gamma(v,u) \cap C| \geq |\Gamma(v,u)|/6$, which allows translating quality guarantees that are quantified in terms of $|\Gamma(v,u)|$ to similar guarantees in terms of $|C'_{v,u}|$. In fact, $|C'_{v,u} \setminus (\bigcup_{j \leq i-1} L_j)| \geq |\Gamma(v,u)|/6$, which enables further translation of these guarantees to quantification in terms of $|C'_{v,u} \cap R_{i-1}|$.

Claim 5.2.4 *Referring to the partition reconstruction procedure, for every $i \in [\ell]$, the following holds.*

1. *The number of missing edges inside the bi-cliques formed in iteration i is at most $12\gamma_2\epsilon \cdot N^2$; that is,*
$$\left| \bigcup_{(v,u) \in F_i} \{(w_1,w_2) \in \Gamma'(v) \times \Gamma'(u) : (w_1,w_2) \notin E\} \right| \leq 12\gamma_2\epsilon \cdot N^2.$$

2. *The number of* ("superfluous") *edges inside the bi-cliques formed in iteration i is at most $12\gamma_2\epsilon \cdot N^2$; that is,*
$$\left| \bigcup_{(v,u) \in F_i} \{(w_1,w_2) \in (\Gamma'(v) \times \Gamma'(v)) \cup (\Gamma'(u) \times \Gamma'(u)) : (w_1,w_2) \in E\} \right| \leq 12\gamma_2\epsilon \cdot N^2.$$

3. *The number of* ("superfluous") *edges between bi-cliques formed in iteration i and either R_i or other bi-cliques formed in the same iteration is at most $36\ell \cdot \gamma_2\epsilon \cdot N^2$; actually,*
$$\left| \bigcup_{(v,u) \in F_i} \{(w_1,w_2) \in C'_{v,u} \times (R_{i-1} \setminus C'_{v,u}) : (u,w) \in E\} \right| \leq 36\ell \cdot \gamma_2\epsilon \cdot N^2.$$

4. *$|R_i| \leq 2^{-i} \cdot N$ and $|L_i| \leq 2^{-(i-1)} \cdot N$.*

Thus, the total number of violations caused by the bi-cliques that are formed by the foregoing procedure is upper-bounded by $(36+o(1))\ell^2 \cdot \gamma_2\epsilon \cdot N^2 = o(\epsilon N^2)$.

Proof: We prove all items simultaneously, by induction from $i = 0$ to $i = \ell$. Needless to say, all items hold vacuously for $i = 0$, and thus we focus on the induction step.

Starting with Item 1, we note that every $(v,u) \in F_i$ is i-good and thus the number of edges missing in $\Gamma'(v) \times \Gamma'(u) \subseteq \Gamma(v) \times \Gamma(u)$ is at most $\gamma_2 2^i \epsilon \cdot |\Gamma(v,u)| \cdot N$. As in the proof of Claim 3.2.4, we need to relate $|\Gamma(v,u)|$ to $|C'_{v,u} \cap R_{i-1}|$ (in order to upper-bound the contribution of all pairs in F_i). We recall that $C'_{v,u} = C_{v,u} \setminus C$, where C is the set of vertices that are already covered when this bi-clique $\Gamma(v,u)$ is identified. Also recall that $|\Gamma(v,u) \setminus C_{v,u}| = o(1) \cdot |\Gamma(v,u)|$ and

$|(\Gamma(v,u) \setminus C) \setminus L| \geq |\Gamma(v,u)|/5$, where $L \stackrel{\rm def}{=} \bigcup_{j\in[i-1]} L_j$. Using $C'_{v,u} = (C'_{v,u} \cap R_{i-1}) \cup (C'_{v,u} \cap L)$, we get that $C'_{v,u} \cap R_{i-1} = (C_{v,u} \setminus C) \setminus L$ and it follows that $|C'_{v,u} \cap R_{i-1}| \geq |(\Gamma(v,u) \setminus C) \setminus L| - o(|\Gamma(v,u)|) > |\Gamma(v,u)|/6$. Combining all the above (and recalling that the sets $C'_{v,u}$ are disjoint), we obtain

$$\begin{aligned} \left| \bigcup_{(v,u)\in F_i} \{(w_1,w_2) \in \Gamma'(v) \times \Gamma'(u) : (w_1,w_2) \notin E\} \right| &\leq \gamma_2 2^i \epsilon \cdot \sum_{(v,u)\in F_i} |\Gamma(v,u)| \cdot N \\ &\leq \gamma_2 2^i \epsilon \cdot 6|R_{i-1}| \cdot N. \end{aligned}$$

Using the induction hypothesis regarding R_{i-1} (i.e., $|R_{i-1}| < 2^{-(i-1)} \cdot N$), Item 1 follows.

Item 2 is proved in a similar fashion. As for Item 3, we adapt the proof of Item 2 of Claim 3.2.4. Specifically, the number of edges in $C_{v,u} \times ([N] \setminus C_{v,u})$ is upper-bounded by the sum of $|C_{v,u} \times (\Gamma(v,u) \setminus C_{v,u})|$ and the number of edges in $C_{v,u} \times ([N] \setminus \Gamma(v,u))$. Using Condition 3 of i-goodness (of (v,u)), we upper-bound both $|\Gamma(v,u) \setminus C_{v,u}|$ and the number of edges of the second type. Hence, the number of edges in $C'_{v,u} \times (R_{i-1} \setminus C'_{v,u}) \subseteq C_{v,u} \times ([N] \setminus C_{v,u})$ is at most $3\ell \cdot \gamma_2 2^i \epsilon \cdot |\Gamma(v,u)| \cdot N$. Using again $\sum_{(v,u)\in F_i} |\Gamma(v,u)| < 6|R_{i-1}|$ and $|R_{i-1}| < 2^{-(i-1)} \cdot N$, we establish Item 3.

Turning to Item 4, we first note that $L_i \subseteq R_{i-1}$ and thus $|L_i| \leq |R_{i-1}| \leq 2^{-(i-1)} \cdot N$. As for R_i, let us consider all the cases that might lead to placing a vertex v in R_i; that is, the various violations of the three conditions in Step 1.

Violation of Condition (b): not having sufficiently high degree. We observe that vertices that violate Condition (b) do not contribute to R_i, because each such vertex is either covered in iteration i or ends-up in L_i. Specifically, let v be an arbitrary vertex that violates Condition (b), and let $k(v) \in \{0, 1, ..., \ell'\}$ be the largest index k such that $\phi_k(v)$ holds (where ϕ_0 is fictitiously defined such that it always holds). Then, Condition (b) is equivalent to requiring that $|\Gamma(v)| \geq 0.9^{k(v)} \cdot \beta_1 \cdot 2^i \epsilon \cdot N$ holds. Indeed, if the latter condition does not hold, then v is placed in L_i (and the converse holds as well).

In the subsequent cases, we shall assume that Condition (b) does hold with respect to the vertex v.

Violation of Condition (a): not being i-good. Here we refer to vertices that are not i-good although they have degree at least $\beta_2 \cdot 2^i \epsilon \cdot N > \gamma_2 \cdot 2^i \epsilon \cdot N$. By Claim 5.2.3, the number of vertices of this type is at most $\gamma_1 2^{-i} \cdot N$.

Violation of Condition (c). Here we refer to vertices that satisfy both Conditions (a) and (b) but violate Condition (c), which refers to the existence of a good edge that yields a bi-clique with sufficiently many new vertices. The rest of the proof is devoted to upper-bounding the number of such vertices. Loosely speaking, this is done by using the upper bound established in Item 3, while relying on the hypothesis that these vertices satisfy both Conditions (a) and (b).

Recalling that we refer to vertices that satisfy both Conditions (a) and (b), we first upper-bound the number of vertices that have relatively many neighbors in the current C (i.e., vertices v such that $|\Gamma(v) \cap C| \geq |\Gamma(v)|/8$). As in the proof of Claim 3.2.4, each such vertex v requires at least $|\Gamma(v)|/8 \geq \beta_2 \cdot 2^i \epsilon \cdot N/8$ edges from $C' \stackrel{\rm def}{=} \bigcup_{(v',u')\in \bigcup_{j\in[i]} F_j} C'_{v',u'}$ to it, whereas by Item 3 the total number of edges going out from C' to R_i is at most $i \cdot 36\ell \cdot \gamma_2 \epsilon \cdot N^2$. Hence, the number of vertices of this type is upper-bounded by

$$\frac{36\ell^2 \cdot \gamma_2 \epsilon \cdot N^2}{\beta_2 \cdot 2^i \epsilon \cdot N} = \frac{36\ell^2 \cdot \gamma_2}{\beta_2} \cdot 2^{-i} N < 0.1 \cdot 2^{-i} N, \tag{12}$$

where the last inequality uses $\gamma_2 < \beta_2/(360\ell^2)$.

In the rest of the proof we consider only vertices that have have relatively few neighbors in the current C (i.e., $|\Gamma(v) \cap C| \leq |\Gamma(v)|/8$). In particular, by the case hypothesis (i.e., v is i-good), there exist $u \notin C$ such that (v,u) is i-good (because the fraction of "non-good" pairs is at most $\gamma_3 < 1/2$). Thus, we focus on the condition $|(\Gamma(v,u) \setminus C) \setminus L| > |\Gamma(v,u)|/5$, where $L \stackrel{\rm def}{=} \bigcup_{j \leq i-1} L_j$ and C denotes the current set of covered vertices. We distinguish three cases with respect to the relation between $|\Gamma(v)|$ and $|\Gamma(u)|$.

Case of $|\Gamma(v)| \gg |\Gamma(u)|$ (i.e., $|\Gamma(v)| > 1.3|\Gamma(u)|$). Using the case hypothesis (which implies $|\Gamma(v)| > |\Gamma(v,u)|/2$), it suffices to show that $|(\Gamma(v) \setminus C) \setminus L| > |\Gamma(v)|/2$. Since $|\Gamma(v) \cap C| \leq |\Gamma(v)|/8$, we focus on upper-bounding $|\Gamma(v) \cap L|$ for typical v. The intuition is that in the current case $\neg\phi_1(v)$ holds and so ($v \notin L_i$ implies) $|\Gamma(v)| \geq \beta_1 \cdot 2^i \epsilon N$, whereas each vertex in $\Gamma(v) \cap L_j$ has at most $\beta_2 \cdot 2^j \epsilon N$ neighbors of degree at least $\beta_1 \cdot 2^i \epsilon N$ (which yields a total count of $2\beta_2 \epsilon N^2$ edges in $L_j \times (R_{i-1} \setminus L_i)$). Thus, the number of vertices $v \in R_{i-1} \setminus L_i$ for which $|\Gamma(v) \cap L| > |\Gamma(v)|/8$ holds is sufficiently small. Details follow.

Using the hypothesis that (v,u) is i-good (and referring to Condition 2 of Definition 5.2.1), we note that the number of edges with both endpoints in $\Gamma(v)$ is at most $\gamma_2 \cdot 2^i \epsilon \cdot |\Gamma(v,u)| \cdot N \leq \gamma_2 \cdot 2^{i+1} \epsilon \cdot |\Gamma(v)| \cdot N$. Thus, less than $(200\ell)^{-1}$ fraction of the vertices in $\Gamma(v)$ have more that $200\ell \cdot \gamma_2 \cdot 2^{i+1} \epsilon \cdot N < \beta_2 \cdot 2^i \epsilon \cdot N/100 \leq |\Gamma(v)|/100$ such edges, where the inequalities are due to $\gamma_2 \leq \beta_2/40000\ell$ and $|\Gamma(v)| \geq \beta_2 \cdot 2^i \epsilon \cdot N$ (since $v \notin L_i$). By Condition 3 of Definition 5.2.1, at most $(200\ell)^{-1}$ fraction of the vertices in $\Gamma(v)$ have at least $200\ell \cdot \gamma_2 \cdot 2^i \epsilon \cdot N < |\Gamma(v)|/100$ edges going out of $\Gamma(v,u)$. We conclude that less than a $(100\ell)^{-1}$ fraction of the vertices in $\Gamma(v)$ have degree exceeding $|\Gamma(u)| + 0.02|\Gamma(v)| < |\Gamma(v)|$, and so $\neg\phi_1(v)$ holds. The latter fact allows us to increase our lower-bound on $|\Gamma(v)|$ (from $|\Gamma(v)| \geq \beta_2 \cdot 2^i \epsilon N$) to $|\Gamma(v)| \geq \beta_1 \cdot 2^i \epsilon N$ (using again $v \notin L_i$). Thus, if $|\Gamma(v) \cap L| > |\Gamma(v)|/8$ then there exist at least $\beta_1 \cdot 2^i \epsilon N/8$ edges from $L = \bigcup_{j \leq i-1} L_j$ to v.

We upper-bound the number of such vertices v (i.e., for which $|\Gamma(v) \cap L| > |\Gamma(v)|/8$), by upper-bounding the number of edges that may go from L to any vertex of degree at least $\beta_1 \cdot 2^i \epsilon N$. The contribution of each vertex in $L_j^{(2)}$ to this number is at most $\beta_2 \cdot 2^j \epsilon N$, because vertices in $L_j^{(2)}$ have degree at most $\beta_2 \cdot 2^j \epsilon N$. As for the vertices in $L_j \setminus L_j^{(2)}$, each such vertex u' violates $\phi_{\ell'}$ and thus can contribute at most $|\Gamma(u')|/100\ell$ to this number, because at most a $1/100\ell$ fraction of its neighbors have degree exceeding $1.2|\Gamma(u')| < \beta_1 \cdot 2^i \epsilon N$ (since $|\Gamma(u')| < \beta_1 \cdot 2^j \epsilon N$ and $j \leq i-1$), whereas we count edges to vertices of degree at least $\beta_1 \cdot 2^i \epsilon N$. Thus, the contribution of each vertex in $u' \in L_j$ to the count is at most $\max(\beta_2 \cdot 2^j \epsilon N, |\Gamma(u')|/100\ell) \leq \beta_1 \cdot 2^j \epsilon N/100\ell$ (since $\beta_2 \leq \beta_1/100\ell$ and $|\Gamma(u')| < \beta_1 \cdot 2^j \epsilon N$). Recalling that $|L_j| \leq |R_{j-1}| \leq 2^{-(j-1)} N$, it follows that the number of bad vertices (i.e., vertices v of degree at least $\beta_1 \cdot 2^i \epsilon N$ with at least $|\Gamma(v)|/8$ neighbors in L) is at most

$$\begin{aligned} \frac{\sum_{j \leq i-1} |L_j| \cdot \beta_1 \cdot 2^j \epsilon \cdot N/100\ell}{\beta_1 \cdot 2^i \epsilon N/8} &\leq \frac{(i-1) \cdot \beta_1 \cdot 2\epsilon \cdot N^2/100\ell}{\beta_1 \cdot 2^i \epsilon N/8} \\ &< 0.16 \cdot 2^{-i} N, \end{aligned}$$

whereas the rest of the vertices $v \in R_{i-1} \setminus L_i$ satisfy $|\Gamma(v) \cap L| \leq |\Gamma(v)|/8$. Recalling that $|\Gamma(v) \cap C| \leq |\Gamma(v)|/8$, we conclude that $|(\Gamma(v) \setminus C) \setminus L| > |\Gamma(v)|/2$, and the claim follows; that is, the current case is only responsible for $0.16 \cdot 2^{-i}N$ vertices violating Condition (c).

Case of $|\Gamma(v)| \ll |\Gamma(u)|$ (i.e., $|\Gamma(v)| < 0.7|\Gamma(u)|$). In this case we shall show that $|(\Gamma(u) \setminus C) \setminus L| > |\Gamma(u)|/2$ (and use $|\Gamma(u)| > |\Gamma(v,u)|/2$). We first show that $|\Gamma(u) \cap L| \leq |\Gamma(u)|/8$, and later turn to show that typically $|\Gamma(u) \cap C| \leq |\Gamma(u)|/8$ holds as well. The proof of the first claim is supported by the intuition that almost all vertices in $\Gamma(u)$ have the approximately the same degree as v and satisfy $\phi_{\ell'}$ (since most of their neighbors have degree approximately $|\Gamma(u)| \gg |\Gamma(v)|$), which implies that they cannot be in L (because vertices in L that satisfy $\phi_{\ell'}$ have degree at most $\beta_2 \cdot 2^{i-1}\epsilon N$, whereas $v \in R_{i-1} \setminus L_i$ has degree at least $\beta_2 \cdot 2^i \epsilon N$). Details follow.

We start by showing that almost all vertices in $\Gamma(u)$ satisfy $\phi_{\ell'}$. Analogously to the previous case, at most 1% of the vertices in $\Gamma(u)$ have more than $0.02 \cdot |\Gamma(v)|$ neighbors not in $\Gamma(v)$. On the other hand, by using Condition 1 of Definition 5.2.1, at least 99% of the vertices in $\Gamma(u)$ have at least $0.99 \cdot |\Gamma(v)|$ neighbors in $\Gamma(v)$, whereas at least 99% of the vertices in $\Gamma(v)$ have degree at least $0.99 \cdot |\Gamma(u)|$. Let us denote by V the subset of $\Gamma(u)$ containing vertices v' such that $|\Gamma(v')| \leq 1.02 \cdot |\Gamma(v)|$ and $\Gamma(v') \cap \Gamma(v)$ contains at least $0.98 \cdot |\Gamma(v)|$ vertices of degree at least $0.99 \cdot |\Gamma(u)|$. Then, $|V| > 0.98|\Gamma(u)|$, because 98% of the vertices in $\Gamma(u)$ have both degree at most $1.02 \cdot |\Gamma(v)|$ and at least $0.99 \cdot |\Gamma(v)|$ neighbors in $\Gamma(v)$ (whereas at most 1% of the vertices in $\Gamma(v)$ have degree smaller than $0.99 \cdot |\Gamma(u)|$). We note that each vertex in V has degree at most $1.02 \cdot |\Gamma(v)| < 0.72 \cdot |\Gamma(u)|$, whereas at least a $0.98/1.02 \gg (100\ell)^{-1}$ fraction of its neighbors have degree at least $0.99 \cdot |\Gamma(u)| > 1.2 \cdot 0.72 \cdot |\Gamma(u)|$, which implies that each vertex in V satisfies $\phi_{\ell'}$. Using the latter fact and recalling that each vertex in V has degree at least $0.99 \cdot |\Gamma(v)| \geq 0.99 \cdot \beta_2 \cdot 2^i \epsilon N$ (since $v \notin L_i$), we show that $V \cap L = \emptyset$. The latter claim follows by noting that for every $v' \in L$ that satisfies $\phi_{\ell'}$ it holds that $|\Gamma(v')| < \beta_2 \cdot 2^{i-1}\epsilon N$, whereas every $v' \in V$ satisfies both $\phi_{\ell'}$ and $|\Gamma(v')| > 0.99 \cdot \beta_2 \cdot 2^i \epsilon N$. Finally, using $V \cap L = \emptyset$ and $|V| \geq 0.98|\Gamma(u)|$, we get $|\Gamma(u) \cap L| \leq |\Gamma(u) \setminus V| \leq 0.02|\Gamma(u)|$.

Having established $|\Gamma(u) \cap L| \leq |\Gamma(u)|/8$, we now turn to provide a similar upper-bound for $|\Gamma(u) \cap C|$. Unlike in the previous case (or rather in the preliminary proof that $\Gamma(v) \cap C$ is small), here we cannot *directly* charge the vertices in $\Gamma(u) \cap C$ to edges going out from C to v. Still an indirect charging rule will work; that is, we first charge such vertices to u, and then distribute the charge to u's neighbors.

Specifically, suppose that $|\Gamma(u) \cap C| > |\Gamma(u)|/8$. This means that there are at least $|\Gamma(u)|/8$ edges going out from C to u. Wishing to charge these edges to the initial vertex v (while considering all initial $v \in R_{i-1} \setminus L_i$), we charge each neighbor of u by one eighth of an edge (i.e., 1/8 unit) as its share in the edges going from C to u. (This guarantees that, when considering different initial vertices, it still holds that each edge going out of C is charged at most 1 unit.) Indeed, an important observation is that we are not concerned with the existence of a specific $u \in \Gamma(v)$ that violates $|\Gamma(u) \cap C| \leq |\Gamma(u)|/8$, but should be concerned only if this violation occurs for all $u \in \Gamma(v) \setminus C$ such that (v, u) is i-good (and $|\Gamma(u)| > |\Gamma(v)|/0.7$), since otherwise we may just pick some $u \in \Gamma(v) \setminus C$ such that (v, u) is i-good and $|\Gamma(u) \cap C| \leq |\Gamma(u)|/8$. Thus, we get into trouble with v only if, for every $u \in \Gamma(v) \setminus C$ that (v, u) is i-good, both $|\Gamma(u)| > |\Gamma(v)|/0.7$ and $|\Gamma(u) \cap C| > |\Gamma(u)|/8$ hold.[15] Let us denote the set of such bad vertices by B, and note that each vertex $v \in B$ is charged with at least $(|\Gamma(v)|/2) \cdot (1/8) > \beta_2 \cdot 2^i \epsilon N/16$

[15] If $|\Gamma(u)| > |\Gamma(v)|/0.7$ does not hold then this u is handled in the other two cases.

edges going from C to $\Gamma(v)$, where $\Gamma(v)|/2$ is a lower-bound on the number of vertices $u \in \Gamma(v)$ such that $u \notin C$ and (v,u) is i-good.[16] Since the total number of edges going out from C is at most $36\ell^2 \cdot \gamma_2 \epsilon \cdot N^2$, we upper-bound $|B|$ by $0.1 \cdot 2^{-i}N$ (as in Eq. (12), except that here we use $\gamma_2 < \beta_2/(6000\ell^2)$). To re-cap, note that we showed that the current case is only responsible for $0.1 \cdot 2^{-i}N$ vertices that violating Condition (c).

Case of $|\Gamma(v)| \approx |\Gamma(u)|$ (i.e., $0.7|\Gamma(u)| \leq |\Gamma(v)| \leq 1.3|\Gamma(u)|$. We first note that the analysis of $|\Gamma(u) \cap C|$ for a typical (v,u), as presented in the previous case (of $|\Gamma(v)| \ll |\Gamma(u)|$), still applies. Thus, for all but $0.1 \cdot 2^{-i}N$ vertices v, there exists a vertex u such that either the first case holds (i.e., $|\Gamma(v)| > 1.3|\Gamma(u)|$) or $|\Gamma(u) \cap C| \leq |\Gamma(u)|/8$. (If the first case holds then we proceeds as in the first case, and otherwise we proceed as follows.) We shall show, below, that $|\Gamma(u) \cap L| \leq |\Gamma(u)|/8$, and conclude that $|(\Gamma(u) \setminus C) \setminus L| \geq |\Gamma(u)|/2$, which in turn is lower-bounded by $|\Gamma(v,u)|/5$ (since $|\Gamma(u)| \geq |\Gamma(v,u)|/2.3$).

The claim $|\Gamma(u) \cap L| \leq |\Gamma(u)|/8$ is supported by the intuition that almost all vertices in $\Gamma(u)$ have approximately the same degree as v. However, in the current case these vertices do not necessarily satisfy $\phi_{\ell'}$ and so their being in L does not necessarily mean their having degree below $\beta_2 \cdot 2^{i-1}\epsilon N$, which is significantly smaller than $|\Gamma(v)| \geq \beta_2 \cdot 2^i \epsilon N$. So we need a different method to argue that being in L is inconsistent with having degree approximately $|\Gamma(v)|$. Indeed, the source of trouble is that for two different thresholds $\beta' > \beta''$ it may be the case that $v \notin L_i$ holds because $|\Gamma(v)| \geq \beta'' \cdot 2^i \epsilon N$, whereas $v' \in L_j$ holds because $|\Gamma(v')| < \beta' \cdot 2^j \epsilon N$. Here is where the intermediate thresholds (and the different ϕ_k) come into play: we shall show that whenever the foregoing happens it holds that $\beta' \approx \beta''$ (rather than $\beta' > 2\beta''$, which would have not given anything). Specifically, we shall show that if $\phi_k(v)$ holds then $\phi_{k-1}(v')$ must hold for almost all $v' \in \Gamma(u)$. Thus, if $v \notin L_i$ due to $|\Gamma(v)| \geq 0.9^k \beta_1 \cdot 2^i \epsilon N$ (and $\phi_k(v)$ holds), then $v' \in L_j$ implies that $|\Gamma(v')| < 0.9^{k-1}\beta_1 \cdot 2^j \epsilon N$, which yields the desired contradiction. Details follow.

Using arguments as in the previous two cases, we first establish that at least 99% of the vertices in $\Gamma(u)$ have degree at most $(1+\ell^{-2}) \cdot |\Gamma(v)|$ and have at least $(1-\ell^{-2}) \cdot |\Gamma(v)|$ neighbors in $\Gamma(v)$. (Here the argument relies on $\gamma_2 \leq \beta_2/(500\ell^2)$ and $|\Gamma(u)| \geq |\Gamma(v)|/1.3 \geq \beta_2 \cdot 2^i \epsilon N/1.3$.) Let us denote this (large) subset of $\Gamma(u)$ by V, and note that $v \in V$. Similarly, one can show that at least $1-(200\ell)^{-1}$ of the vertices in $\Gamma(v)$ have degrees in the interval $[(1 \pm (300\ell')^{-1}) \cdot |\Gamma(u)|]$. Hence, for every $v' \in V$, it holds that $|\Gamma(v')|$ is in the interval $(1 \pm (300\ell')^{-1}) \cdot |\Gamma(v)|$, whereas at least $\frac{1-(200\ell)^{-1}}{1+\ell^{-2}} > 1-(100\ell)^{-1}$ of its neighbors (i.e., the vertices in $\Gamma(v')$) have degrees in the interval $[(1 \pm (300\ell')^{-1}) \cdot |\Gamma(u)|]$. Denoting (for every $v' \in V$),

$$\rho(v') \stackrel{\text{def}}{=} \max_{S \subseteq \Gamma(v') \text{ s.t. } |S|=|\Gamma(v')|/100\ell} \left\{ \min_{u' \in S} \left\{ \frac{|\Gamma(u')|}{|\Gamma(v')|} \right\} \right\} \tag{13}$$

we infer that for every $v' \in V$ (including v) it holds that $\rho(v') = \frac{(1\pm(300\ell')^{-1}) \cdot |\Gamma(u)|}{(1\pm(300\ell')^{-1}) \cdot |\Gamma(v)|} = (1 \pm (100\ell')^{-1}) \cdot \frac{|\Gamma(u)|}{|\Gamma(v)|}$. It follows that $\rho(v') \geq \frac{1-(100\ell')^{-1}}{1+(100\ell')^{-1}} \cdot \rho(v) > (1-(30\ell')^{-1}) \cdot \rho(v)$.

Recall that $k(v') \in \{0,1,...,\ell'\}$ is the largest index k such that $\phi_k(v')$ holds (where ϕ_0 always holds). Indeed, $\rho(v) > 1.1 + \frac{k(v)}{10\ell'}$ and $|\Gamma(v)| \geq 0.9^{k(v)} \cdot \beta_1 \cdot 2^i \epsilon \cdot N$ (because $v \notin L_i$). Combining $\rho(v') > (1-(30\ell')^{-1}) \cdot \rho(v)$ and $\rho(v) > 1.1 + \frac{k(v)}{10\ell'}$, it follows that for every $v' \in V$ it holds that

[16] Recall that the fraction of vertices $u \in \Gamma(v)$ such that $u \in C$ is at most 1/8, whereas the fraction of vertices $u \in \Gamma(v)$ such that (v,u) is not i-good is $\gamma_3 < 3/8$.

$\rho(v') > 1.1 + \frac{k(v)-1}{10\ell'}$, which implies $k(v') \geq k(v)-1$. It follows that $V \cap L = \emptyset$, because otherwise we obtain, for some $j \leq i-1$, a vertex $v' \in V \cap L_j$ such that $|\Gamma(v')| < 0.9^{k(v')} \cdot \beta_1 \cdot 2^j \epsilon \cdot N \leq 0.9^{k(v)-1} \cdot \beta_1 \cdot 2^{i-1}\epsilon \cdot N \leq |\Gamma(v)|/1.8$, which contradicts $|\Gamma(v')| \geq (1-(300\ell')^{-1}) \cdot |\Gamma(v)| > \Gamma(v)|/1.8$. Recalling that $|V| \geq 0.99 \cdot |\Gamma(u)|$, we conclude that $|\Gamma(u) \cap L| \leq 0.01|\Gamma(u)|$.

Combining the preliminary bound (of Eq. (12)) and the bounds of the foregoing three cases, we conclude that at most $(0.1+0.16+0.1+0.1) \cdot 2^{-i}N < 0.5 \cdot 2^{-i}N$ vertices satisfy conditions (a) and (b) but violate Condition (c).

Recall that R_i only contains vertices that satisfy Condition (b) but violate either Condition (a) or Condition (c). The number of the former was upper-bounded by $\gamma_1 \cdot 2^{-i}N$, whereas the number of the latter was just upper-bounded by $0.5 \cdot 2^{-i}N$. Thus, $|R_i| \leq (\gamma_1 + 0.5) \cdot 2^{-i} \cdot N$, and Item 4 follows by the foregoing setting of $\gamma_1 \leq 1/2$. This completes the proof of the current claim. □

Completing the reconstruction and its analysis. The foregoing construction leaves "unassigned" the vertices in R_ℓ as well as some of the vertices in $L_1, ..., L_\ell$. (Note that some vertices in $\bigcup_{i=1}^{\ell-1} L_i$ may be placed in bi-cliques constructed in later iterations, but there is no guarantee that this actually happens.) For sake of elegance, we assign each of these remaining vertices to a two-vertex bi-clique (i.e., an isolated pair of vertices connected by an edge). Ignoring the number of edges used in these bi-cliques (which is negligible), the number of violation caused by this assignment equals the number of edges with both endpoints in $R' \stackrel{\text{def}}{=} R_\ell \cup (\bigcup_{i=1}^{\ell} L_i)$, because edges with a single endpoint in R' were already accounted for in Item 3 of Claim 5.2.4. Nevertheless, we upper-bound the number of violations by the total number of edges incident to R', which in turn is upper-bounded by

$$\begin{aligned}
\sum_{v \in R_\ell \cup (\bigcup_{i\in[\ell]} L_i)} |\Gamma(v)| &\leq |R_\ell| \cdot N + \sum_{i=1}^{\ell} \sum_{v \in L_i} |\Gamma(v)| \\
&\leq \frac{\epsilon N}{4} \cdot N + \sum_{i=1}^{\ell} 2^{-(i-1)} N \cdot \beta_1 2^i \epsilon N \\
&= \frac{\epsilon}{4} \cdot N^2 + 2\ell \cdot \beta_1 \cdot \epsilon N^2.
\end{aligned}$$

By the foregoing setting of β_1 (i.e., $\beta_1 \leq 1/4\ell$), it follows that the number of these edges is smaller than $\epsilon N^2/2$. Combining this with the bounds on the number of violating edges (or non-edges) as provided by Claim 5.2.4, the lemma follows. ■

5.2 Non-Adaptive Lower-Bound for Bi-Clique Collection

In this section we establish Part 2 of Theorem 1.2 by adapting the proof presented in Section 4.1. Specifically, for every value of $\epsilon > 0$, we consider two different classes of graphs, one consisting of graphs in $\mathcal{BCC}$ and the other consisting of graphs that are ϵ-far from $\mathcal{BCC}$, and show that a non-adaptive algorithm of query complexity $o(\epsilon^{-3/2})$ cannot distinguish between graphs selected at random in these classes.

The first class, denoted $\mathcal{BCC}_\epsilon$, contains all N-vertex graphs such that each graph consists of $(16\epsilon)^{-1}$ bi-cliques, and each bi-clique has $8\epsilon \cdot N$ vertices on each side. It will be instructive to partition these $(16\epsilon)^{-1}$ bi-cliques into $(32\epsilon)^{-1}$ pairs (each consisting of two bi-cliques), and view each of these bi-cliques as a super-cycle of length four with $4\epsilon \cdot N$ vertices in each of its four independent sets. The second class, denoted $\mathcal{SC}_8\mathcal{C}_\epsilon$, contains all N-vertex graphs such that each graph consists of $(32\epsilon)^{-1}$ super-cycles of length 8, and each of these super-cycles has $4\epsilon \cdot N$ vertices in each of its eight independent sets. Indeed, $\mathcal{BCC}_\epsilon \subseteq \mathcal{BCC}$, whereas each graph in $\mathcal{SC}_8\mathcal{C}_\epsilon$ is ϵ-far from $\mathcal{BCC}$ (because each of the super-cycles of length 8 must be turned into a collection of bi-cliques). We note that *both classes contain only bipartite graphs.*

In order to motivate the claim that a non-adaptive algorithm of query complexity $o(\epsilon^{-3/2})$ cannot distinguish between graphs selected at random in these classes, consider the algorithm that selects $o(\epsilon^{-3/4})$ vertices and inspects the induced subgraph. Consider the partition of a graph in $\mathcal{SC}_8\mathcal{C}_\epsilon$ into $(32\epsilon)^{-1}$ pairs of bi-cliques (equiv., super-cycles of length 4), and correspondingly the partition of a graph in $\mathcal{SC}_8\mathcal{C}_\epsilon$ into $(32\epsilon)^{-1}$ super-cycles of length 8. Then, the probability that a sample of $o(\epsilon^{-3/4})$ vertices contains at least four vertices that reside in the same part (of $32\epsilon \cdot N$ vertices) is $o(\epsilon^{-3/4})^4 \cdot (32\epsilon)^3 = o(1)$. On the other hand, one may show that if this event does not occur, then the answers obtained from both graphs are indistinguishable. As will be shown below, this intuition extends to an arbitrary non-adaptive algorithm.

As in Section 4.1, it suffices to consider deterministic algorithms. We shall show that, for every set of $o(\epsilon^{-3/2})$ queries, the answers provided by a randomly selected element of $\mathcal{BCC}_\epsilon$ are statistically close to the answers provided by a randomly selected element of $\mathcal{SC}_8\mathcal{C}_\epsilon$. As in Section 4.1, for an N-vertex graph G and a query (u, v), we denote the corresponding answer by $\text{ans}_G(u, v)$.

Lemma 5.3 *Let G_1 and G_2 be random N-vertex graphs uniformly distributed in $\mathcal{BCC}_\epsilon$ and $\mathcal{SC}_8\mathcal{C}_\epsilon$, respectively. Then, for every sequence $(v_1, v_2), ..., (v_{2q-1}, v_{2q}) \in [N] \times [N]$, where the v_i's are not necessarily distinct, it holds that the statistical difference between $\text{ans}_{G_1}(v_1, v_2), ..., \text{ans}_{G_1}(v_{2q-1}, v_{2q})$ and $\text{ans}_{G_2}(v_1, v_2), ..., \text{ans}_{G_2}(v_{2q-1}, v_{2q})$ is $O(q^2\epsilon^3)$.*

Part 2 of Theorem 1.2 follows.

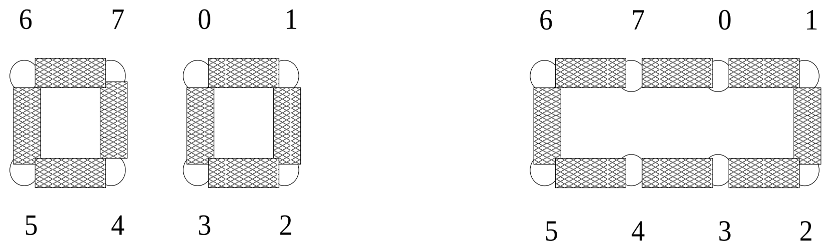

Figure 6: A single part, consisting of eight independent sets, in $\mathcal{BCC}_\epsilon$ and $\mathcal{SC}_8\mathcal{C}_\epsilon$.

Proof: We adapt the proof of Lemma 4.1. Here, we consider a 1-1 correspondence, denoted ϕ, between the vertices of an N-vertex graph in $\mathcal{BCC}_\epsilon \cup \mathcal{SC}_8\mathcal{C}_\epsilon$ and triples in $[(32\epsilon)^{-1}] \times \{0, 1, ..., 7\} \times [4\epsilon \cdot N]$. Specifically, $\phi(v) = (i, j, w)$ indicates that v resides in the $(j+1)^{\text{st}}$ independent set of the i^{th} part of the graph, and it is vertex number w in this set. Recall that in the case of a graph in $\mathcal{BCC}_\epsilon$ the eight independent sets are arranged in two super-paths (each of length 4), whereas in the case of a graph in $\mathcal{SC}_8\mathcal{C}_\epsilon$ the eight independent sets are arranged in a single super-path of length 8. (See Figure 6.) Consequently, the answers provided by uniformly distributed $G_1 \in \mathcal{BCC}_\epsilon$ and $G_2 \in \mathcal{SC}_8\mathcal{C}_\epsilon$ can be emulated by the following two corresponding random processes.

1. The process A_1 selects uniformly a bijection $\phi : [N] \rightarrow [(32\epsilon)^{-1}] \times \{0, 1, ..., 7\} \times [4\epsilon \cdot N]$ and answers each query $(u, v) \in [N] \times [N]$ by 1 if and only if for $\phi(u) = (i_1, j_1, w_1)$ and $\phi(v) = (i_2, j_2, w_2)$ it holds that both $i_1 = i_2$ and $j_1 = (j_2 \pm 1 \bmod 4) + \lfloor j_2/4 \rfloor \cdot 4$.

2. The process A_2 selects uniformly a bijection $\phi : [N] \to [(32\epsilon)^{-1}] \times \{0,1,...,7\} \times [4\epsilon \cdot N]$ and answers each query $(u,v) \in [N] \times [N]$ by 1 if and only if for $\phi(u) = (i_1, j_1, w_1)$ and $\phi(v) = (i_2, j_2, w_2)$ it holds that both $i_1 = i_2$ and $j_1 = j_2 \pm 1 \bmod 8$.

Let us denote by $\phi'(v)$ (resp., $\phi''(v)$ and $\phi'''(v)$) the first (resp., second and third) coordinates of $\phi(v)$; that is, $\phi(v) = (\phi'(v), \phi''(v), \phi'''(v))$. Then, both processes answer the query (u,v) with 0 if $\phi'(u) \neq \phi'(v)$, and the difference between the processes is confined to the case that $\phi'(u) = \phi'(v)$. Specifically, conditioned on $\phi'(u) = \phi'(v)$, it holds that $A_1(u,v) = 1$ if and only if $\phi''(u) = (\phi''(v) \pm 1 \bmod 4) + \lfloor \phi''(v)/4 \rfloor \cdot 4$, whereas $A_2(u,v) = 1$ if and only if $\phi''(u) = \phi''(v) \pm 1 \bmod 8$. However, since the (random) value of ϕ'' is not present at the answer, the foregoing difference may go unnoticed. These considerations apply to a single query, but things may change in case of several queries. In general, the event that allows distinguishing the two processes is a simple cycle of at least four vertices that have the same ϕ' value. Minor differences may also be due to equal ϕ''' values, and so we also consider these in our "bad" event.

Definition 5.3.1 *We say that ϕ is* bad (w.r.t the sequence $(v_1, v_2), ..., (v_{2q-1}, v_{2q}) \in [N] \times [N]$), *if one of the following two conditions hold:*

1. *For some $i \in [(32\epsilon)^{-1}]$, the subgraph $Q_i = (V_i, E_i)$, where $V_i = \{v_k : k \in [2q] \wedge \phi'(v) = i\}$ and $E_i = \{\{v_{2k-1}, v_{2k}\} : v_{2k-1}, v_{2k} \in V_i\}$, contains a simple cycle of length at least four.*
2. *There exists $i \neq j \in [2q]$ such that $\phi'''(v_i) = \phi'''(v_j)$.*

Indeed, the query sequence $(v_1, v_2), ..., (v_{2q-1}, v_{2q})$ will be fixed throughout the rest of the proof, and so we shall omit it from our terminology.

Claim 5.3.2 *The probability that a uniformly distributed bijection ϕ is bad is at most*

$$O(q^2\epsilon^3) + \frac{q^2}{16\epsilon N}$$

Proof: We start by upper-bounding the probability that the second event in Definition 5.3.1 holds. We have $\binom{2q}{2}$ sub-events, and each holds with probability $1/(32\epsilon \cdot N)$. As for the first event, for every $t \geq 4$, we upper-bound the probability that some Q_i contains a simple cycle of length t. As in the proof of Claim 4.1.2, we observe that the query graph contains at most $(2q)^{t/2}$ cycles of length t, whereas the probability that a specific simple t-cycle is contained in some Q_i is $(32\epsilon)^{t-1}$. Thus, the probability of the first event is upper-bounded by

$$\sum_{t \geq 4} (2q)^{t/2} \cdot (32\epsilon)^{t-1} < \sum_{t \geq 4} \left(\sqrt{2q} \cdot 32 \cdot \epsilon^{(t-1)/t}\right)^t < \sum_{t \geq 4} \left(50\sqrt{q} \cdot \epsilon^{3/4}\right)^t,$$

which is upper-bounded by $2 \cdot (50\sqrt{q} \cdot \epsilon^{3/4})^4 = O(q^2\epsilon^3)$, provided that $50\sqrt{q} \cdot \epsilon^{3/4} < 1/2$ (and the claim hold trivially otherwise). □

Claim 5.3.3 *Conditioned on the bijection ϕ not being bad, the sequences $(A_1(v_1, v_2), ..., A_1(v_{2q-1}, v_{2q}))$ and $(A_2(v_1, v_2), ..., A_2(v_{2q-1}, v_{2q}))$ are identically distributed.*

Proof: Noting that Definition 5.3.1 only refers to ϕ' and ϕ''', we fixed any choice of ϕ' and ϕ''' that yields a good ϕ and consider the residual random choice of ϕ''. Referring to the foregoing subgraphs

Q_i's, recall that pairs with endpoints in different Q_i's are answered by 0 in both processes. Note that (by the second condition in Definition 5.3.1) the hypothesis implies that ϕ''' assigns different values to the different vertices in $\{v_k : k \in [2q]\}$, and it follows that ϕ'' assigns these vertices values that are uniformly and independently distributed in $\{0, 1, ..., 7\}$. Now, using the first condition in Definition 5.3.1, the hypothesis implies that the only simple cycles appearing in $Q_i = (V_i, E_i)$ have length three. We shall show that this implies that (in each of the two processes) the answer assigned to each edge in Q_i is independent of the answer given to other edges of Q_i.

We first note that, in each of the two processes, every query (v_{2k-1}, v_{2k}) such that $\phi''(v_{2k-1}) \equiv \phi''(v_{2k-1}) \pmod 2$ is answered negatively (i.e., in such a case, $A_1(v_{2k-1}, v_{2k}) = A_2(v_{2k-1}, v_{2k}) = 0$). Thus, fixing any (random) values of $(\phi''(v_k) \bmod 2 : k \in [2q])$, we may omit from $Q_i = (V_i, E_i)$ all edges that connect vertices that have the same value of $\phi'' \pmod 2$, because the answers to these queries are already determined (as 0, in each of the two processes). This omission eliminates (from Q_i) all cycles of length three, which are the only simple cycles in the original Q_i, and thus each modified Q_i is a forest. We can now proceed analogously to the proof of Claim 4.1.3, although things are slightly more complex here. Specifically, we consider the residual random values of ϕ'' (conditioned on $\phi'' \bmod 2$); that is, we augments the fixed values of $\phi'' \bmod 2$ with the random values of $\lfloor \phi''/2 \rfloor$, which are uniformly distributed in $\{0, 1, 2, 3\}$. We view these random selections as taking place in an order determined by some fixed traversal of each tree (of the aforementioned forest), and note that at each step (and in each of the processes) the new random value (uniformly distributed in $\{0, 1, 2, 3\}$) yields answer 1 (to the corresponding query) with probability 1/2.

1. In the case of A_1, the query/edge $(u, v) \in E_i$ (which satisfies $\phi'(u) = i = \phi'(v)$ and $\phi''(u) \equiv \phi''(v) + 1 \pmod 2$) is answered 1 if and only if $\phi''(u) = (\phi''(v) \pm 1 \bmod 4) + \lfloor \phi''(v)/4 \rfloor \cdot 4$ holds (which means that $\lfloor \phi''(u)/4 \rfloor = \lfloor \phi''(v)/4 \rfloor$). Thus, $A_1(u, v) = 1$ with probability 1/2.

2. In the case of A_2, the query/edge $(u, v) \in E_i$ (which satisfies $\phi'(u) = i = \phi'(v)$ and $\phi''(u) \equiv \phi''(v) + 1 \pmod 2$) is answered 1 if and only if $\phi''(u) = \phi''(v) \pm 1 \bmod 8$ holds. Thus, $A_2(u, v) = 1$ with probability 2/4.

Thus, in each of the two processes, each query is answered by the value 1 with probability exactly 1/2, independently of the answers to all other queries. The claim follows. □

Combining Claims 5.3.2 and 5.3.3, it follows that the statistical distance between the sequences $(A_1(v_1, v_2), ..., A_1(v_{2q-1}, v_{2q}))$ and $(A_2(v_1, v_2), ..., A_2(v_{2q-1}, v_{2q}))$ is at most $O(q^2\epsilon^3 + q^2(\epsilon N)^{-1})$, and the lemma follows for sufficiently large N. ■

5.3 Non-Adaptive Lower-Bound for Super-Cycle Collection

In this section we establish a lower-bound on the non-adaptive query complexity of testing Super-Cycle Collections. We do so by generalizing the ideas presented in Section 5.2.

Specifically, fixing any $t \geq 4$, for every value of $\epsilon > 0$, we consider two different classes of graphs, one consisting of graphs in $\mathcal{S}\mathtt{C}_t\mathcal{C}$ and the other consisting of graphs that are ϵ-far from $\mathcal{S}\mathtt{C}_t\mathcal{C}$, and show that a non-adaptive algorithm of query complexity $o(\epsilon^{-(2t-2)/t})$ cannot distinguish between graphs selected at random in these classes.

The first class, denoted $\mathcal{S}\mathtt{C}_t\mathcal{C}_\epsilon$, contains all N-vertex graphs such that each graph consists of $(t^2\epsilon)^{-1}$ super-cycles of length t, and each super-cycle has $t\epsilon \cdot N$ vertices in each of its t independent sets. It will be instructive to partition these $(t^2\epsilon)^{-1}$ super-cycles into $(2t^2\epsilon)^{-1}$ pairs. The second class, denoted $\mathcal{S}\mathtt{C}_{2t}\mathcal{C}_\epsilon$, contains all N-vertex graphs such that each graph consists of $(2t^2\epsilon)^{-1}$ super-cycles of length $2t$, and each super-cycle has $t\epsilon \cdot N$ vertices in each of its $2t$ independent sets. Indeed, $\mathcal{S}\mathtt{C}_t\mathcal{C}_\epsilon \subseteq \mathcal{S}\mathtt{C}_t\mathcal{C}$, whereas each graph in $\mathcal{S}\mathtt{C}_{2t}\mathcal{C}_\epsilon$ is ϵ-far from $\mathcal{S}\mathtt{C}_t\mathcal{C}$ (because each of the super-cycles of length $2t$ must be turned into a pair of super-cycles of length t).

As in Section 5.2, we motivate the claim that a non-adaptive algorithm of query complexity $o(\epsilon^{-(2t-2)/t})$ cannot distinguish between graphs selected at random in these classes by considering a specific algorithm that inspects the subgraph induced by a random set of $o(\epsilon^{-(t-1)/t})$ vertices. The probability that a sample of $o(\epsilon^{-(t-1)/t})$ vertices contains at least t vertices that reside in the same part (of $(2t^2\epsilon)\cdot N$ vertices) is $\binom{o(\epsilon^{-(t-1)/t})}{t} \cdot (2t^2\epsilon)^{t-1} = o(1)$, where the o-notation refers to a fixed value of t and a varying value of $\epsilon > 0$. On the other hand, one may show that if this event does not occur, then the answers obtained from both graphs are indistinguishable. As will be shown below, this intuition extends to an arbitrary non-adaptive algorithm. Following the same conventions as in Section 5.2, it suffices to prove the following

Lemma 5.4 (Lemma 5.3, generalized): *For every fixed $t \geq 4$, let G_1 and G_2 be random N-vertex graphs uniformly distributed in $\mathcal{S}\mathtt{C}_t\mathcal{C}_\epsilon$ and $\mathcal{S}\mathtt{C}_{2t}\mathcal{C}_\epsilon$, respectively. Then, for every sequence $(v_1, v_2), ..., (v_{2q-1}, v_{2q}) \in [N] \times [N]$, where the v_i's are not necessarily distinct, it holds that the statistical difference between $\mathrm{ans}_{G_1}(v_1, v_2), ..., \mathrm{ans}_{G_1}(v_{2q-1}, v_{2q})$ and $\mathrm{ans}_{G_2}(v_1, v_2), ..., \mathrm{ans}_{G_2}(v_{2q-1}, v_{2q})$ is $O(q^{t/2}\epsilon^{t-1})$.*

Part 2 of Conjecture 1.3 follows. Indeed, Lemma 5.3 is obtained as a special case (of Lemma 5.4) by setting $t = 4$. The following proof is slightly different from the proof provided in Section 5.2.

Proof: We generalize the proof of Lemma 5.3. We consider a bijection, denoted ϕ, between the vertices of an N-vertex graph in $\mathcal{S}\mathtt{C}_t\mathcal{C}_\epsilon \cup \mathcal{S}\mathtt{C}_{2t}\mathcal{C}_\epsilon$ and triples in $[(2t^2\epsilon)^{-1}] \times \{0, 1, ..., 2t-1\} \times [t\epsilon \cdot N]$. Specifically, $\phi(v) = (i, j, w)$ indicates that v resides in the $(j+1)^{\text{st}}$ independent set of the i^{th} part of the graph, and that it is vertex number w in this set. Recall that in the case of a graph in $\mathcal{S}\mathtt{C}_t\mathcal{C}_\epsilon$ the $2t$ independent sets in each part are arranged in two super-paths (each of length t), whereas in the case of a graph in $\mathcal{S}\mathtt{C}_{2t}\mathcal{C}_\epsilon$ the $2t$ independent sets are arranged in a single super-path of length $2t$. Consequently, the answers provided by uniformly distributed $G_1 \in \mathcal{S}\mathtt{C}_t\mathcal{C}_\epsilon$ and $G_2 \in \mathcal{S}\mathtt{C}_{2t}\mathcal{C}_\epsilon$ can be emulated by the following two corresponding random processes.

1. The process A_1 selects uniformly a bijection $\phi : [N] \to [(2t^2\epsilon)^{-1}] \times \{0, 1, ..., 2t-1\} \times [t\epsilon \cdot N]$ and answers each query $(u, v) \in [N] \times [N]$ by 1 if and only if for $\phi(u) = (i_1, j_1, w_1)$ and $\phi(v) = (i_2, j_2, w_2)$ it holds that both $i_1 = i_2$ and $j_1 = (j_2 \pm 1 \bmod t) + \lfloor j_2/t \rfloor \cdot t$.
2. The process A_2 selects uniformly a bijection $\phi : [N] \to [(2t^2\epsilon)^{-1}] \times \{0, 1, ..., 2t-1\} \times [t\epsilon \cdot N]$ and answers each query $(u, v) \in [N] \times [N]$ by 1 if and only if for $\phi(u) = (i_1, j_1, w_1)$ and $\phi(v) = (i_2, j_2, w_2)$ it holds that both $i_1 = i_2$ and $j_1 = j_2 \pm 1 \bmod 2t$.

Again, let us denote by $\phi'(v)$ (resp., $\phi''(v)$ and $\phi'''(v)$) the first (resp., second and third) coordinates of $\phi(v)$; that is, $\phi(v) = (\phi'(v), \phi''(v), \phi'''(v))$. Then, both processes answer the query (u, v) with 0 if $\phi'(u) \neq \phi'(v)$, and the difference between the processes is confined to the case that $\phi'(u) = \phi'(v)$. Specifically, conditioned on $\phi'(u) = \phi'(v)$, it holds that $A_1(u, v) = 1$ if and only if $\phi''(u) = (\phi''(v) \pm 1 \bmod t) + \lfloor \phi''(v)/t \rfloor \cdot t$, whereas $A_2(u, v) = 1$ if and only if $\phi''(u) = \phi''(v) \pm 1 \bmod 2t$. In general, the event that allows distinguishing the two processes is a simple cycle of at least t vertices that have the same ϕ' value. Minor differences may also be due to equal ϕ''' values, and so we also consider these in our "bad" event.

Definition 5.4.1 (Definition 5.3.1, generalized): *We say that ϕ is* bad (w.r.t the sequence of queries $(v_1, v_2), ..., (v_{2q-1}, v_{2q}) \in [N] \times [N]$), *if one of the following two conditions hold:*

1. *For some $i \in [(2t^2\epsilon)^{-1}]$, the subgraph $Q_i = (V_i, E_i)$, where $V_i = \{v_k : k \in [2q] \wedge \phi'(v) = i\}$ and $E_i = \{\{v_{2k-1}, v_{2k}\} : v_{2k-1}, v_{2k} \in V_i\}$, contains a simple cycle of length at least t.*

2. *There exists $i \neq j \in [2q]$ such that $\phi'''(v_i) = \phi'''(v_j)$.*

Indeed, the query sequence $(v_1, v_2), ..., (v_{2q-1}, v_{2q})$ will be fixed throughout the rest of the proof, and so we shall omit it from our terminology.

Claim 5.4.2 (Claim 5.3.2, generalized): *The probability that a uniformly distributed bijection ϕ is bad is at most*

$$O(t)^{2t} \cdot q^{t/2} \epsilon^{t-1} + \frac{q^2}{t^2 \epsilon N}$$

Proof: We start by upper-bounding the probability that the second event in Definition 5.4.1 holds. We have $\binom{2q}{2}$ sub-events, and each holds with probability $1/(2t^2\epsilon \cdot N)$. As for the first event, for every $\ell \geq t$, we upper-bound the probability that some Q_i contains a simple cycle of length ℓ by $(2q)^{\ell/2} \cdot (2t^2\epsilon)^{\ell-1}$. Thus, the probability of the first event is upper-bounded by

$$\sum_{\ell \geq t} (2q)^{\ell/2} \cdot (2t^2\epsilon)^{\ell-1} \;<\; \sum_{\ell \geq t} \left(3t^2\sqrt{q} \cdot \epsilon^{(t-1)/t}\right)^{\ell},$$

which is upper-bounded by $2 \cdot (3t^2\sqrt{q} \cdot \epsilon^{(t-1)/t})^t = O(t)^{2t} \cdot q^{t/2}\epsilon^{t-1}$, provided that $3t^2\sqrt{q} \cdot \epsilon^{(t-1)/t} < 1/2$ (and the claim hold trivially otherwise). □

Claim 5.4.3 (Claim 5.3.3, generalized): *Conditioned on the bijection ϕ not being bad, the sequences $(A_1(v_1, v_2), ..., A_1(v_{2q-1}, v_{2q}))$ and $(A_2(v_1, v_2), ..., A_2(v_{2q-1}, v_{2q}))$ are identically distributed.*

Proving this claim is the only difficulty in extending the proof of Lemma 5.3 to the current setting. Indeed, the following proof yields a slightly different proof of Claim 5.3.3.

Proof: Again, we fix any choice of ϕ' and ϕ''' that yields a good ϕ, and consider the residual random choice of $\phi''(v_1), ..., \phi''(v_{2q})$, which (by the second hypothesis in Definition 5.4.1) are uniformly and independently distributed in $\{0, 1, ..., 2t-1\}$. Considering any of the aforementioned graphs $Q_i = (V_i, E_i)$, we note that this graph does not contain simple cycles of length greater than $t-1$.

We now consider $\phi'' : V_i \to \{0, 1, ..., 2t-1\}$ as being selected at random in two stages. In the first stage we assign each vertex a random value mod t, and in the second stage we assign each vertex a random bit representing its most significant bit; that is, for each vertex $v \in V_i$, we first determine (at random) the value $\phi''(v) \bmod t$, which we denote by $\psi''(v)$, and next determine (at random) the bit $\lfloor \phi''(v)/t \rfloor$, which we denote by $\pi''(v)$. Thus, $\phi''(v) = \psi''(v) + \pi''(v) \cdot t$, and it will be instructive to depict the graphs as in Figure 7. Fixing an arbitrary setting of values for the first stage, we shall consider what may happen in the second stage.

For every fixed setting of ψ'', we consider the residual graph $Q'_i = (V_i, E'_i)$, where E'_i contains only the queries in E_i that are still undetermined (given ψ''); that is, $(u, v) \in E_i$ is placed in E'_i if and only if $\psi''(u) \equiv \psi''(v) \pm 1 \pmod t$, whereas all the other queries (or rather the answers to them) are already determined (as being answered by 0). We shall consider the connected components of Q'_i, and show that (conditioned on the foregoing setting of ψ'') the answers provided to the queries

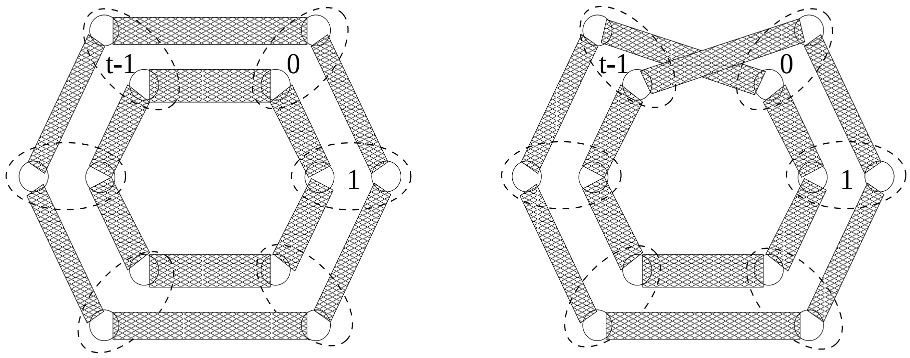

Figure 7: A single part, consisting of $2t$ independent sets, in $\mathcal{S}\mathtt{C}_t\mathcal{C}_\epsilon$ and $\mathcal{S}\mathtt{C}_{2t}\mathcal{C}_\epsilon$. The ellipses indicate the values of ψ''.

in E_i' under A_1 are distributed identically to the answers provided under A_2. Specifically, for each possible sequence of answers, we shall show a 1-1 correspondence between the assignments of π'' that yield these answers under A_1 and the assignments of π'' that yield these answers under A_2. (Recall that $\phi''(v) = \psi''(v) + \pi''(v) \cdot t$.) That is, for each possible sequence of answers and each connected component of Q_i', we shall show that the number of assignments of π'' that yield these answers under A_j is independent of $j \in \{1, 2\}$.

Let $C = (V_i'', E_i'')$ be an arbitrary connected component of $Q_i' = (V_i, E')$, and let $A'' : E_i'' \to \{0, 1\}$ describe an arbitrary sequence of answers to the queries E_i''. Our aim is proving that the number of assignments of π'' that yield these answers under A_j (i.e., satisfy $A_j(u, w) = A''(u, w)$ for every $(u, w) \in E_i''$) is independent of $j \in \{1, 2\}$. Furthermore, we shall show that this number is either two or zero (when considering only the assignment of π'' to V_i''). Consider any spanning tree T of C, rooted at an arbitrary vertex $v \in V_i''$. For each choice of $\sigma \in \{0, 1\}$, we shall prove that there exists a unique assignment $\pi'' : V_i'' \to \{0, 1\}$ such that $\pi''(v) = \sigma$ and π'' is consistent with A'' and A_1 (resp., A_2) on the edges of T. That is, the resulting π'' is such that the answers as mandated by A'' for the edges of T fit the answers that A_1 (resp., A_2) provides with respect to $\phi'' = \psi'' + t \cdot \pi''$. As we shall see, these assignments might be inconsistent with the value of A'' on edges that do not belong to the spanning tree. However, we shall show that there is an inconsistency when fitting A_1 if and only if there is an inconsistency when fitting A_2. Details follow.

Fitting the process A_1: Recall that the value of π'' on the root of T was set to σ. The value of π'' on all other vertices is set, by traversing the tree T, in the following manner. When traversing the tree edge (u, w) from a vertex u for which $\pi''(u)$ was already determined to a new w (for which $\pi''(w)$ is still undetermined), we set $\pi''(w) \leftarrow \pi''(u)$ if $A''(u, w) = 1$ and $\pi''(w) \leftarrow 1 - \pi''(u)$ otherwise (i.e., if $A''(u, w) = 0$).

Note that this process determines the values of the bits $\pi''(w)$ for all $w \in V_i''$ such that the tree-neighbors u and w are assigned the same bit if and only if $A''(u, w) = 1$. This is indeed consistent with the definition of A_1. Furthermore, the setting of the values of π'' is uniquely determined by the requirement to be consistent with A_1.

Fitting the process A_2: We assign values exactly as in the case of fitting A_1, with a single exception that refers to the case that the tree-edge $(u, w) \in E_i''$ satisfies $\{\psi''(u), \psi''(w)\} = \{0, t - 1\}$. In this case (where vertex u has already been assigned a value), we set $\pi''(w) \leftarrow 1 - \pi''(u)$ if $A''(u, w) = 1$ and $\pi''(w) \leftarrow \pi''(u)$ otherwise (i.e., if $A''(u, w) = 0$).

That is, in this case (i.e., $\{\psi''(u), \psi''(w)\} = \{0, t-1\}$), the process determines the value of $\pi''(w)$ such that the tree-neighbors u and w are assigned the opposite bits if and only if $A''(u,w) = 1$.

As noted in the foregoing discussion, while each of the two assignments is consistent with A'' (and the corresponding A_j) on the edges of the spanning tree T, there may be inconsistencies with the edges of E_i'' that are not tree edges. It remains to show that there is an inconsistency with respect to the process A_1 if and only if there is an inconsistency with respect to the process A_2.

We shall say that an edge $(u,w) \in E_i''$ (e.g., an edge of the spanning tree T) is a crossing edge if $\{\psi''(u), \psi''(w)\} = \{0, t-1\}$. By definition of the two assignments, the only difference between them is caused when traversing a tree edge that is a crossing edge. For such an edge, the value of π'' is flipped when fitting the process A_2 if and only if it is *not* flipped when fitting the process A_1. Thus, for each $u \in V_i''$, the value assigned to $\pi''(u)$ when fitting A_2 is the XOR of the value assigned to $\pi''(u)$ when fitting A_1 and the *parity* of the number of crossing edges that belong to the tree path from (the root) v to u.

Now, consider an edge $(u,w) \in E_i''$ that is not an edge in the spanning tree T. Consider the simple tree paths from the root v to vertices u and w, respectively, and let us denote their branching point by v'. Let p_u (resp., p_w) be the path on the spanning tree T leading from v' to u (resp., w), and p_u' be the path from v' to u obtained by augmenting p_w with the (non-tree) edge (w,u). Then, the union of p_u and p_u' constitutes a simple cycle, which by the hypothesis has length smaller than t. As we shall show in the next paragraph, it follows that *the parity of the number of crossing edges on p_u equals the parity of the number of crossing edges on p_u'*. In other words, the parity of the number of crossing edges on p_u equals the parity of the number of crossing edges on p_w if and only if (u,w) is not a crossing edge. Assuming that (u,w) is not a crossing edge, consider the value assigned to $\pi''(u)$ and $\pi''(w)$ when fitting A_1 (by following the paths from the root to u and w, respectively). Then, $A''(u,w)$ is inconsistent with $\pi''(u)$ and $\pi''(w)$ as determined when fitting the process A_1 if and only if $A''(u,w)$ is inconsistent with $\pi''(u)$ and $\pi''(w)$ as determined when fitting the process A_2, because in both cases $\pi''(u) \oplus \pi''(w)$ is the same value (since the total number of crossing edges on p_v and p_w is even). A similar argument holds when (u,w) is a crossing edge (since then $\pi''(u) \oplus \pi''(w)$ flips from A_1 to A_2), and the claim follows.

To verify the assertion regarding the parity of the number of crossing edges on p_u and on p_u', consider the values assigned by ψ'' to the vertices in the union of p_u and p_u'. Since the union of p_u and p_u' is a cycle of length less than t, these values must belong to a proper subset, S, of $\{0, ..., t-1\}$. If this set does not contain $\{0, t-1\}$, then we are done (since neither of the paths may contain a crossing edge). Otherwise, for some j, it holds that S is a subset of the union of $S_1 = \{j+1, ..., t-1\}$ and $S_2 = \{0, ..., j-1\}$. If $\psi''(v')$ and $\psi''(u)$ belong to the same S_k, then the parity of the number of crossing edges on both p_u and p_u' is even (since these paths can only move from one subset to the other via a crossing edge).[17] Similarly, if $\psi''(v')$ and $\psi''(u)$ do not belong to the same subset then the parity on each of these paths must be odd. □

Combining Claims 5.4.2 and 5.4.3, the lemma follows. ■

[17]Note that the ψ''-values of intermediate vertices along any path must be "adjacent" modulo t, and so moving between $\{j+1, ..., t-1\}$ and $\{0, ..., j-1\}$ is only possible via $(t-1, 0)$.

5.4 A candidate adaptive tester for Super-Cycle Collection

In this section we outline an adaptive $\widetilde{O}(\epsilon^{-1})$-query algorithm what we conjecture to be a tester for $\mathcal{S}\mathtt{C}_t\mathcal{C}$, where $t \geq 5$ is fixed. The algorithm is a significant generalization of Algorithm 5.1, and we focus on outlining the corresponding sub-test, denoted sub-test$_i(v)$.

Recall that in Algorithm 5.1 this sub-test consists, essentially, of finding an edge (v, u) and checking the potential bi-clique induced by it (i.e., $\Gamma(u) \times \Gamma(v)$). In the current context we try to find a t-cycle $(v_0, v_1, ..., v_{t-1})$ such that $v_0 = v$ and for every $j \in \{0, ..., t-1\}$ it holds that $v_j \in \Gamma(v_{j-1 \bmod t}) \cap \Gamma(v_{j+1 \bmod t}) \neq \Gamma(v_{j-1 \bmod t}) \cup \Gamma(v_{j+1 \bmod t})$. Given such a candidate t-cycle $\overline{v}$, letting $I_j(\overline{v}) \stackrel{\rm def}{=} (\Gamma(v_{j-1 \bmod t}) \cap \Gamma(v_{j+1 \bmod t})$, we check that $I_j(\overline{v}) \times I_{j+1 \bmod t}(\overline{v})$ is a bi-clique, and that $\Gamma(v_j) = I_{j-1 \bmod t}(\overline{v}) \cup I_{j+1 \bmod t}(\overline{v})$. Each of these activities is is to be performed by making $\mathrm{poly}(\log(1/\epsilon))/(2^i\epsilon)$ queries. The implementation of the various checks is similar to the implementation of similar checks performed in Algorithm 5.1, and so we focus on finding the aforementioned t-cycle.

Starting with $v_0 \stackrel{\rm def}{=} v$, we obtain $v_1 \in \Gamma(v)$ just as (u was obtained) in Algorithm 5.1. In fact, we may obtain $v_{t-1} \in \Gamma(v)$ in the same way, except that we need to verify that the latter vertex is actually in a different independent set than v_1. This is done by checking that $\Gamma(v_{t-1})$ is different from $\Gamma(v_1)$, where any w in the symmetric difference of $\Gamma(v_1)$ and $\Gamma(v_{t-1})$ can serve as a witness. (Indeed, $w \in \Gamma(v_1) \setminus \Gamma(v_{t-1}$ can be used as v_2.) Similarly, when holding a partial path $(v_{t-j}, ..., v_0, ..., v_k)$, we seek a vertex v_{k+1} (resp., $v_{t-(j+1)}$) such that $\Gamma(v_{k+1})$ and $\Gamma(v_{k-1})$ (resp., $\Gamma(v_{t-(j+1)})$ and $\Gamma(v_{t-(j-1)})$) are different. When the path reaches length $t-1$ (i.e., holds t vertices), we treat it as a candidate t-cycle.

We note that, as in the case of Algorithm 5.1, it may happen that the foregoing algorithm fails to find a t-cycle, $(v_0, ..., v_{t-1}$. In this case, the algorithm performs only a subset of the checks outlined above. Specifically, suppose that the algorithm failed to extend the partial path $\overline{v} \stackrel{\rm def}{=} (v_{t-j}, ..., v_0, ..., v_k)$ any further. Then, for intermediate vertices the checks are as before, but for the extremes we should proceed with more care. For example, assuming the path contains at least four vertices, we let $I_{t-j}(\overline{v}) \stackrel{\rm def}{=} (\Gamma(v_{t-j+1 \bmod t}) \setminus I_{t-j+2 \bmod t}(\overline{v})$.

Clearly, the foregoing algorithm always accepts any graph in $\mathcal{S}\mathtt{C}_t\mathcal{C}$. One can also verify that, for every $i \leq \ell \stackrel{\rm def}{=} \log_2(1/\epsilon) + 2$, this algorithm rejects with high probability any graph in $\mathcal{S}\mathtt{C}_{2t}\mathcal{C}_{2^{-i}}$, where $\mathcal{S}\mathtt{C}_{2t}\mathcal{C}_{2^{-i}}$ is as in Lemma 5.4. Since graphs in $\mathcal{S}\mathtt{C}_{2t}\mathcal{C}_{\epsilon/4}$ are ϵ-close to $\mathcal{S}\mathtt{C}_t\mathcal{C}$, we conclude that the aforementioned algorithm distinguishes graphs in $\mathcal{S}\mathtt{C}_t\mathcal{C}$ from graphs in $\mathcal{S}\mathtt{C}_{2t}\mathcal{C}' \stackrel{\rm def}{=} \bigcup_{i \geq 5} \mathcal{S}\mathtt{C}_{2t}\mathcal{C}_{2^{-i}}$ that are ϵ-far from $\mathcal{S}\mathtt{C}_t\mathcal{C}$. This yields an algorithm for testing a promise problem, denoted Π_t, which refers to inputs in $\mathcal{S}\mathtt{C}_t\mathcal{C} \cup \mathcal{S}\mathtt{C}_{2t}\mathcal{C}'$ such that the tester is required to accept inputs in $\mathcal{S}\mathtt{C}_t\mathcal{C}$ and reject inputs (in $\mathcal{S}\mathtt{C}_{2t}\mathcal{C}'$) that are ϵ-far from $\mathcal{S}\mathtt{C}_t\mathcal{C}$.

Theorem 5.5 (an almost-quadratic complexity gap for promise problems): *For every positive integer $t \geq 5$, the promise problem Π_t satisfies the following:*

1. *There exists an adaptive tester of query complexity $\widetilde{O}(\epsilon^{-1})$ for Π_t. Furthermore, this tester runs in time $\widetilde{O}(\epsilon^{-1})$.*
2. *Any non-adaptive tester for Π_t must have query complexity $\Omega(\epsilon^{-2+(2/t)})$.*

Indeed, Part 1 follows by the foregoing algorithm, whereas Part 2 follows from Lemma 5.4. We also note that there exists an efficient *non-adaptive* tester of query complexity $O(\epsilon^{-2+(2/t)})$ for Π_t. This tester merely inspects the subgraph induced by a uniformly selected set of $O(\epsilon^{-1+(1/t)})$ vertices, and rejects if and only if this set contains t vertices such that the subgraph induced by these t vertices is a simple t-vertex path.

6 Non-Adaptive Testing with $\widetilde{O}(1/\epsilon)$ Complexity

We first note that $\Omega(1/\epsilon)$ (adaptive) queries are required for testing any graph property that is non-trivial for testing, where a graph property Π is non-trivial for testing if there exists $\epsilon_0 > 0$ such that for infinitely many $N \in \mathbb{N}$ there exist N-vertex graphs G_1 and G_2 such that $G_1 \in \Pi$ and G_2 is ϵ_0-far from Π. We note that all properties considered in this work are non-trivial for testing. On the other hand, the negation of this (non-triviality) condition means that for every $\epsilon > 0$ and all sufficiently large $N \in \mathbb{N}$ either Π contains no N-vertex graph or all N-vertex graphs are ϵ-close to Π. In such a case (for every such ϵ and N), the tester may decide without even looking at the graph.[18] Turning back to properties that are non-trivial for testing, we prove that any tester for such a property must have query complexity $\Omega(1/\epsilon)$.

Proposition 6.1 *Let Π be a property that is non-trivial for testing. Then, any tester for Π has query complexity $\Omega(1/\epsilon)$.*

Note that the claim holds also for general properties (i.e., arbitrary sets of functions).

Proof: Let $\epsilon_0 > 0$ be as in the definition, and consider any $N \in \mathbb{N}$ such that Π contains some N-vertex graphs as well as some N-vertex graphs that are ϵ-far from Π. Let G_0 be any N-vertex graph that is ϵ-far from Π, let $G_1 \in \Pi$ be an N-vertex graph closest to G_0, and let $\delta > \epsilon$ denote the relative distance between G_0 and G_1. Let D denote the set of vertex pairs on which G_0 and G_1 differ; indeed, $|D| = \delta \cdot N^2$. Now, for every $\epsilon \leq \epsilon_0$, consider a graph, G, obtained at random from G_0 and G_1 by uniformly selecting a random $R \subseteq D$ of cardinality $\epsilon \cdot N^2$ and letting G agree with G_0 on all pairs in R and agree with G_1 otherwise. Clearly, any tester that makes $o(\epsilon_0/\epsilon)$ queries cannot distinguish G from G_1 (because regardless of is query selection strategy, its next query resides in R with probability at most $|R|/|D| \leq \epsilon/\epsilon_0$). Thus, such a tester cannot decide correctly on both G and G_1 (because G is ϵ-far from Π whereas $G_1 \in \Pi$). Recalling that ϵ_0 is a fixed constant, the proposition follows. ■

6.1 Clique and Bi-Clique

We start with the problem of testing whether the given graph is a clique (or, equivalently, an independent set). The algorithm consists of selecting uniformly $O(1/\epsilon)$ vertex-pairs and checking whether each of these pairs is connected by an edge. Clearly, if the graph is ϵ-far from being a clique, then a randomly selected pair of vertices is connected with probability at most $1 - \epsilon$. The foregoing algorithm and analysis seem to provide the simplest example of a graph property that can be tested by $O(1/\epsilon)$ non-adaptive queries. A somewhat less simple example is provided by testing the property of being a bi-clique.

Algorithm 6.2 (non-adaptive test of bi-cliqueness): *On input N and ϵ and oracle access to a graph $G = ([N], E)$, the tester sets $t = \Theta(1/\epsilon)$ and selects arbitrarily a start vertex s* (e.g., $s = 1$).

[18] Indeed, there exists natural graph properties that are trivial for testing (e.g., connectivity, non-planarity, having no vertex of odd degree); see [GGR, Sec. 10.2.1].

For $i = 1, ..., t$, the tester selects uniformly a pair of vertices (u_i, v_i), and makes the queries (s, u_i), (s, v_i), and (u_i, v_i). The tester accepts if and only if for every i an even number of answers are positive (i.e., indicate the existence of an edge).

Clearly, if G is a bi-clique then for every i either all vertices reside on the same side (and so (s, u_i), (s, v_i), and (u_i, v_i) are all non-edges) or a single vertex is in solitude (and is thus adjacent to the other two vertices). To analyze what happens when G is ϵ-far from being a bi-clique we observe that s induces a partition of the graph to neighbors and non-neighbors (i.e., the 2-partition $(\Gamma(s), [N] \setminus \Gamma(s))$). That is, if G were a bi-clique then every vertex $v \in \Gamma(s)$ (resp., $v \in [N] \setminus \Gamma(s)$) would have satisfied $\Gamma(v) = [N] \setminus \Gamma(s)$ (resp., $\Gamma(v) = \Gamma(s)$).[19] However, since G is ϵ-far from being a bi-clique, it follows that either there are $\frac{\epsilon}{2} \cdot N^2$ edges in $(\Gamma(s) \times \Gamma(s)) \cup (([N] \setminus \Gamma(s)) \times ([N] \setminus \Gamma(s)))$ or $\frac{\epsilon}{2} \cdot N^2$ edges are missing from $\Gamma(s) \times ([N] \setminus \Gamma(s))$. Thus, the sample of t pairs will hit such an edge with probability at least $2/3$.

6.2 Collection of a constant number of cliques

For any constant c, we consider the set of graphs that consists of a collection of (up to) c cliques; that is, the property $\mathcal{CC}^{\leq c}$. Note that the special case of $\mathcal{CC}^{\leq 2}$ is analogous to bi-clique, because a graph $G = ([N], E)$ is in $\mathcal{CC}^{\leq 2}$ if and only if its complement graph $([N], ([N] \times [N]) \setminus E)$ is a bi-clique. The general case (i.e., $c \geq 3$) seems less easy (for non-adaptive testers).

Algorithm 6.3 (non-adaptive test for $\mathcal{CC}^{\leq c}$): *On input N and ϵ and oracle access to a graph $G = ([N], E)$, set $\ell = \log_2(1/\epsilon)$ and proceed as follows.*

1. *Select a uniform sample of $\Theta(\epsilon^{-1/2})$ vertices, denoted S, and examine all vertex pairs in S.*
2. *For $i = 1, ..., \ell$ select, uniformly at random, samples of $\Theta(\log(1/\epsilon)/(2^i\epsilon))$ and $\Theta(2^i)$ vertices in $[N]$ denoted T_i^1 and T_i^2, respectively, and a sample of $\Theta(\min\{2^i, 1/(2^i\epsilon)\})$ vertices in S, denoted S_i. Examines all the vertex pairs in $S_i \times (T_i^1 \cup T_i^2)$ and in $T_i^1 \times T_i^2$.*
3. *Accept if and only if the view of the subgraph as obtained in Steps 1-2 is consistent with some graph in $\mathcal{CC}^{\leq c}$. Namely, let $g' : \left((S \times S) \cup \left(\bigcup_{i=1}^{\ell} \left((S_i \times (T_i^1 \cup T_i^2)) \cup (T_i^1 \times T_i^2)\right)\right)\right) \to \{0,1\}$ be the function determined by the answers obtained in Steps 1-2. Then, the test accepts if and only if g' can be extended to a function over $S' \times S'$ that represents a graph in $\mathcal{CC}^{\leq c}$, where $S' \stackrel{\text{def}}{=} S \cup \left(\bigcup_{i=1}^{\ell}(T_i^1 \cup T_i^2)\right)$.*

It is instructive to spell-out the meaning of the acceptance criterion that underlies Step 3. Indeed, this criterion is equivalent to the conjunction of the following four conditions:

(i) The subgraph induced by S is in $\mathcal{CC}^{\leq c}$.

In such a case, we denote the corresponding cliques by $C_1, ..., C_{c'}$, where $c' \leq c$.

(ii) For every $i \in [\ell]$ and every $v \in T_i^1 \cup T_i^2$, either $\Gamma(v) \cap S_i = \emptyset$ or, for some $j \in [c']$, it holds that $\Gamma(v) \cap S_i = C_j \cap S_i$.

(iii) For every $i \in [\ell]$, if $|\{j : C_j \cap S_i \neq \emptyset\}| = c$ then every $v \in T_i^1 \cup T_i^2$ has neighbors in S_i.

[19] Indeed, this is a simple application of the "induced partition" idea, which underlies the analysis of many of the testers of [GGR].

(iv) For every $i \in [\ell]$ and for every $v \in T_i^1$ and $u \in T_i^2$ such that $\Gamma(v) \cap S_i \neq \emptyset$ and $\Gamma(u) \cap S_i \neq \emptyset$ the following holds. If $\Gamma(v) \cap S_i = \Gamma(u) \cap S_i$ then $(v,u) \in E$, while if $\Gamma(v) \cap S_i \neq \Gamma(u) \cap S_i$, then $(v,u) \notin E$.

Algorithm 6.3 has query complexity

$$|S|^2 + \sum_{i=1}^{\ell} \Big(|S_i| \cdot (|T_i^1| + |T_i^2|) + |T_i^1| \cdot |T_i^2| \Big) = O(1/\epsilon) + \log(1/\epsilon) \cdot O(\log(1/\epsilon)/\epsilon) = \widetilde{O}(1/\epsilon)$$

and accepts every graph in $\mathcal{CC}^{\leq c}$ with probability 1. We thus turn to analyze the case that the input graph $G = ([N], E)$ is ϵ-far from $\mathcal{CC}^{\leq c}$. Namely, we show:

Lemma 6.4 *If G is ϵ-far from $\mathcal{CC}^{\leq c}$ then Algorithm 6.3 rejects with probability at least* $2/3$.

Theorem 1.4 follows.

Proof: Consider first the choice of S. We think of S as being selected in $c+1$ phases, where in phase t, a new uniform sample S^t, of $\Theta(\epsilon^{-1/2})$ vertices, is selected (recall that c is a constant). Intuitively, the objective of the first c phases is to ensure, with high (constant) probability, that as long as the number of vertices that do not have any neighbor among the vertices selected so far is relatively big, we obtain such a vertex in the next phase. After c phases we use the selected vertices to define a partition of the graph vertices into at most c subsets with some *exceptional* vertices (which either do not have any neighbor among the vertices selected in the previous phases or are somehow inconsistent with these vertices). The objective of phase $c+1$ is to ensure that (with high probability) the number of exceptional vertices is relatively small (or else, cause rejection). The analysis relies on the fact that $\mathcal{CC}^{\leq c}$ is a hereditary property (i.e., any induced subgraph of any graph in $\mathcal{CC}^{\leq c}$ is also in $\mathcal{CC}^{\leq c}$).

For each $1 \leq t \leq c+1$, let $S^{\leq t} = \bigcup_{k=1}^{t} S^k$. Recall that the algorithm queries all vertex pairs in $S \times S$. Hence, if for any $1 \leq t \leq c+1$, the subgraph induced by $S^{\leq t}$ is not a collection of at most c cliques, then the algorithm rejects, and we are done. Otherwise, let $C_1^t, ..., C_{c^{(t)}}^t$ denote the $c^{(t)} \leq c$ cliques in the subgraph induced by $S^{\leq t}$. For each $1 \leq t \leq c$, we define the following partition of the set $[N]$ of all graph vertices:

$$\begin{aligned}
V_j^t &\stackrel{\rm def}{=} \{v \,:\, \Gamma(v) \cap S^{\leq t} = C_j^t\} \quad \text{for } 1 \leq j \leq c^{(t)}\ , \\
R_0^t &\stackrel{\rm def}{=} \{v \,:\, \Gamma(v) \cap S^{\leq t} = \emptyset\} \\
R_1^t &\stackrel{\rm def}{=} [N] \setminus \Big(R_0^t \cup \Big(\bigcup_{1 \leq j \leq c^{(t)}} V_j \Big) \Big)\ .
\end{aligned}$$

That is, for $1 \leq j \leq c^{(t)}$, the subset V_j^t consists of the vertices that neighbor all vertices in C_j^t and no other vertex in $S^{\leq t}$, the subset R_0^t consists of all vertices that have no neighbor in $S^{\leq t}$, and R_1^t consists of all vertices that either neighbor only some of the vertices in one of the cliques C_j^t (but not all) or have neighbors in more than one of the cliques. Observe that $V_j^{t+1} \subseteq V_j^t$ and $R_0^{t+1} \subseteq R_0^t$ while $R_1^{t+1} \supseteq R_1^t$.

Given the above notation, we make two observations. The first observation is that for any $1 \leq t \leq c$, if S^{t+1} contains some vertex in R_1^t, then the subgraph induced by $S^{\leq (t+1)}$ is not a collection of at most c cliques, and so the algorithm rejects. It follows that if $|R_1^t| > \frac{1}{4}\epsilon^{1/2} N$ for some $t \leq c$, then the algorithm rejects with high probability. The second observation is that if

S^{t+1} contains some vertex in R_0^t, then $c^{(t+1)} \geq c^{(t)} + 1$. Note that, as long as $|R_0^t| > \frac{1}{4}\epsilon^{1/2}N$, the probability that S^{t+1} does not contain any vertex in R_0^t is at a small constant. Therefore, either $|R_0^c| \leq \frac{1}{4}\epsilon^{1/2}N$, or the algorithm rejects with high probability, because the subgraph induced by $S^{\leq(c+1)}$ consists of more than c connected components. From this point on, we assume that the subgraph induced by $S^{\leq(c+1)}$ is a collection of at most c cliques, that $|R_1^c| \leq \frac{1}{4}\epsilon^{1/2}N$ and that $|R_0^c| \leq \frac{1}{4}\epsilon^{1/2}N$. (We later take into account the small constant probability that this is not the case (but that the algorithm did not reject).)

To simplify the notation, we use the shorthand R_0 for R_0^c, and R_1 for R_1^c, the shorthand c' for $c^{(t)}$, and the shorthand V_j for V_j^c. We also denote $R_0 \cup R_1$ by R. We start by making the simplifying assumption that for each sufficiently large V_j, the corresponding C_j contains a number of vertices that is proportional to the size of V_j. To be precise, $|C_j|/|S| \geq \frac{1}{2}(|V_j|/N)$ holds for every $1 \leq j \leq c'$ that satisfies $|V_j| \geq \frac{\epsilon^{-1/2}}{2c}N$. We justify this assumption at the end of the proof.

Recall that G is ϵ-far from $\mathcal{CC}^{\leq c}$. This means that for every partition of the graph vertices into at most c subsets, the total number of vertex pairs that either belong to the same subset but do not have an edge between them, or belong to different subsets but do have an edge between them, is greater than ϵN^2. In particular, this holds for the partition of $[N]$, denoted $(\widetilde{V}_j)_{j\in\{0,1,\ldots,c'\}}$, that we define as follows:

- For every $j \in [c']$, it holds that $V_j \subseteq \widetilde{V}_j$.
- The vertices in R are partitioned among the $\widetilde{V}_j$'s as follows. For every vertex $v \in R$ and $j \in [c']$, let $e_j(v) = |\Gamma(v) \cap V_j|$ (resp., $\bar{e}_j = |V_j \setminus \Gamma(v)|$) be the number of neighbors (resp., non-neighbors) that v has in V_j. If $c' = c$ then each vertex $v \in R$ is placed in the subset $\widetilde{V}_j$ for which $\bar{e}_j(v) + \sum_{k\in[c']\setminus\{j\}} e_k(v)$ is minimized. If $c' < c$ then we do the same, except that every vertex $v \in R$ that satisfies $\sum_{k=1}^{c'} e_k(v) < \min_{j\in[c']}\{\bar{e}_j(v) + \sum_{k\in[c']\setminus\{j\}} e_k(v\}$ is placed in $\widetilde{V}_0$; that is, v is placed in $\widetilde{V}_0$ if for every $j \in [c']$ it holds that $e_j(v) < \bar{e}_j(v)$.

We note that it may be the case that $\widetilde{V}_0 = \emptyset$; indeed, this always happens when $c' = c$.

Recall that $|R| \leq \frac{1}{2}\epsilon^{1/2}N$. Therefore, the total number of vertex pairs in $R \times R$ is at most $\frac{1}{4}\epsilon N^2$. It follows that if G is ϵ-far from $\mathcal{CC}^{\leq c}$ then (at least) one of the following three events must occur:

1. There are at least $\frac{1}{4}\epsilon N^2$ missing edges between pairs of vertices that belong to the same subset V_j; that is, $\sum_{j=1}^{c'} |(V_j \times V_j) \setminus E| \geq \frac{\epsilon}{4}N^2$.
2. There are at least $\frac{1}{4}\epsilon N^2$ superfluous edges between pairs of vertices that belong to different subsets V_j and V_k; that is, $\sum_{j=1}^{c'-1}\sum_{k=j+1}^{c'} |(V_j \times V_k) \cap E| \geq \frac{\epsilon}{4}N^2$.
3. The total number of missing and superfluous edges contributed by pairs of vertices in $R \times (\bigcup_{j=1}^{c'} V_j)$ is at least $\frac{1}{4}\epsilon N^2$. That is, if for each $j \in [c']$ and $v \in R \cap \widetilde{V}_j$ we let
$$x(v) = \bar{e}_j(v) + \sum_{k\in[c']\setminus\{j\}} e_k(v) \;, \tag{14}$$
and for $v \in R \cap \widetilde{V}_0$ we let
$$x(v) = \sum_{1\leq k\leq c'} e_k(v) \;, \tag{15}$$
then $\sum_{j=0}^{c'}\sum_{v\in R\cap\widetilde{V}_j} x(v) \geq \frac{\epsilon}{4}N^2$. (Recall that $\widetilde{V}_0 = \emptyset$ whenever $c' = c$.)

It remains to prove that in each of the three foregoing cases the algorithm rejects with probability at least 5/6. Specifically, we shall show that, with probability at least 5/6, there exists an $i \in [\ell]$ such that the sample $S_i \cup T_i^1 \cup T_i^2$ contains a set of vertices that induce a subgraph not in $\mathcal{CC}^{\leq c}$ that is inspected by the algorithm. More specifically, this set will contain at most one vertex from each T_i^b, and we shall use the fact that the algorithm inspects all pairs in $(S_i \times (T_i^1 \cup T_i^2)) \cup (T_i^1 \times T_i^2) \cup (S_i \times S_i)$. In what follows let $\epsilon' = \frac{\epsilon}{8\ell c^2}$.

Case 1: $\sum_{j=1}^{c'} |(V_j \times V_j) \setminus E| \geq \frac{\epsilon}{4} N^2$. In this case there must be an index $1 \leq j^* \leq c'$ such that the number of missing edges with both endpoints in V_{j^*} is at least $\frac{\epsilon}{4c} N^2$; that is,

$$\sum_{v \in V_{j^*}} |V_{j^*} \setminus (\{v\} \cup \Gamma(v))| \geq \frac{\epsilon}{4c} N^2 \ . \tag{16}$$

In particular this implies that $|V_{j^*}| \geq \frac{\epsilon^{1/2}}{2c^{1/2}} N$. For each $i \in [\ell]$, we define a subset $B_{j^*,i}$ of V_{j^*} as follows.

$$B_{j^*,i} = \left\{ v \in V_{j^*} : \ |V_{j^*} \setminus (\{v\} \cup \Gamma(v))| \geq \frac{N}{2^i} \right\}, \tag{17}$$

where $B_{j^*,0} = \emptyset$. By Eq. (16), we have

$$\sum_{i=1}^{\ell} |B_{j^*,i} \setminus B_{j^*,i-1}| \cdot \frac{N}{2^i} \geq \frac{\epsilon}{4c} N^2 \tag{18}$$

and thus there exists $i^* \in [\ell]$ (i.e., a set B_{j^*,i^*}) such that

$$|B_{j^*,i^*}| \geq \frac{2^{i^*}\epsilon}{4c\ell} N \geq 2^{i^*} \epsilon' N \ . \tag{19}$$

By the definition of $B_{j^*,i}$ if $B_{j^*,i} \neq \emptyset$, then $|V_{j^*}| \geq N/2^{i^*}$. Since $B_{j^*,i^*} \neq \emptyset$, it holds that $|V_{j^*}| \geq \alpha N$ where $\alpha = \max\{1/2^{i^*}, \frac{\epsilon^{1/2}}{2c^{1/2}}\}$. We shall show that, with high probability, the following three events occur: (1) S_{i^*} contains at least one vertex w from C_{j^*}; (2) $T_{i^*}^1$ contains at least one vertex v from $B_{j^*,i^*} \subseteq V_{j^*}$; and (3) $T_{i^*}^2$ contains at least one vertex u from $V_{j^*} \setminus \Gamma(v)$. If the three event occur then the algorithm rejects since it obtains evidence that the graph is not in $\mathcal{CC}^{\leq c}$ (in the form of $(w,v), (w,u) \in E$ and $(v,u) \notin E$). (Indeed, $v \in \Gamma(w)$ since $w \in C_{j^*}$ and $v \in V_{j^*}$, and $u \in \Gamma(w) \setminus \Gamma(v)$ since $u \in V_{j^*} \setminus \Gamma(v)$. Also note that the algorithm queries all pairs in $(S_{i^*} \times (T_{i^*}^1 \cup T_{i^*}^2)) \cup (T_{i^*}^1 \times T_{i^*}^2)$.)

Let α be as defined in the foregoing discussion. Since $|V_{j^*}| \geq \alpha N$ and we assume that $|C_{j^*}|/|S| \geq \frac{1}{2}|V_{j^*}|/N$, the probability that the first event does not occur is at most $(1-\alpha/2)^{|S_{i^*}|}$ which is a small constant (due to our choice of $|S_{i^*}| = \Theta(1/\alpha)$). Similarly (by our choice of $|T_{i^*}^1| = \Theta(\log(1/\epsilon)/(\epsilon 2^{i^*})) = \Theta(\ell/(\epsilon 2^{i^*})) = \Omega(1/(\epsilon' 2^{i^*})))$, the probability that $T_{i^*}^1$ does not contain any vertex from B_{j^*,i^*} is a small constant (due to the density of B_{j^*,i^*} as lower-bounded in Eq. (19)). Finally, assuming that $T_{i^*}^1$ contains a vertex $v \in B_{j^*,i^*}$, the probability that $T_{i^*}^2$ (which has size $\Theta(2^{i^*})$) does not contain any vertex from $V_{j^*} \setminus \Gamma(v)$ is a small constant as well (since, by definition of B_{j^*,i^*}, the set $V_{j^*} \setminus \Gamma(v)$ has density at least 2^{-i^*}).

Case 2: $\sum_{j=1}^{c'-1} \sum_{k=j+1}^{c'} |(V_j \times V_k) \cap E| \geq \frac{\epsilon}{4} N^2$. In this case there exists at least one pair of subsets, V_{j^*} and V_{k^*} (where $j^* \neq k^*$), such that $|(V_{j^*} \times V_{k^*}) \cap E| \geq \frac{\epsilon}{4c^2} N^2$. Assume, without loss of generality, that $|V_{j^*}| \geq |V_{k^*}|$, so that in particular $|V_{j^*}| \geq \frac{\epsilon^{1/2}}{2c} N$. Similarly to Case 1, it follows that there exists a index $i^* \in \{1, \ldots, \ell\}$ and a subset $B_{j^*,i^*} \subseteq V_{j^*}$ such that $|B_{j^*,i^*}| \geq \epsilon' 2^{i^*} N$ and for every $v \in B_{j^*,i^*}$ it holds that $|V_{k^*} \cap \Gamma(v)| \geq N/2^{i^*}$. Analogously to Case 1, here we can show that, with high probability, the following three events occur: (1) S_{i^*} contains at least one vertex w from C_{j^*}, (2) $T^1_{i^*}$ contains at least one vertex v from B_{j^*,i^*}, and (3) $T^2_{i^*}$ contains at least one vertex u from $V_{k^*} \cap \Gamma(v)$. If these three events occur then the algorithm rejects since it obtains evidence that the graph is not in $\mathcal{CC}^{\leq c}$ (in the form of $(w,v) \in E$, $(w,u) \notin E$ and $(v,u) \in E$). The probability that these three events occur is lower-bounded as in Case 1.

Case 3: $\sum_{j=0}^{c'} \sum_{v \in R \cap \widetilde{V}_j} x(v) \geq \frac{\epsilon}{4} N^2$. For each $v \in R$, let $x(v)$ be as defined in Eq. (14) & (15), and let $R' \stackrel{\rm def}{=} \left\{ v \in R \,:\, x(v) \geq \frac{\epsilon^{1/2}}{4} N \right\}$. Since $|R| \leq \frac{1}{2}\epsilon^{1/2}N$, we have that $\sum_{j=0}^{c} \sum_{v \in (R \setminus R') \cap V_j} x(v) < |R| \cdot \frac{\epsilon^{1/2}}{4} N \leq \frac{\epsilon}{8} N^2$. Therefore, $\sum_{j=0}^{c} \sum_{v \in R' \cap V_j} x(v) \geq \frac{\epsilon}{8} N^2$. By the definition of R', for every $v \in R'$, we have that $x(v) \geq N/2^i$ for some $i \leq \ell/2+2$. Therefore, if we define $B_i = \{v \,:\, x(v) \geq N/2^i\}$ for $i = 1, \ldots, \ell/2+2$, then there is an index $i^* \in [\ell/2+2]$ such that $|B_{i^*}| \geq \frac{\epsilon}{8\ell} 2^{i^*} N > \epsilon' 2^{i^*} N$. Similarly to the previous cases, with high probability, the sample $T^1_{i^*}$ contains at least one vertex v in B_{i^*}. We next show that for each fixed choice of such a vertex $v \in B_{i^*}$, with high probability over the choice of the samples S_{i^*} and $T^2_{i^*}$, we obtain evidence containing v that G is not in $\mathcal{CC}^{\leq c}$ (i.e., a set of vertices that induce a subgraph not in $\mathcal{CC}^{\leq c}$, while having at most one vertex in each $T^b_{i^*}$).

Let $j^* \in \{0, 1, ..., c'\}$ be such that $v \in \widetilde{V}_{j^*}$, and define $\bar{e}_0(v) = e_0(v) = 0$. Observe that since $v \in \widetilde{V}_{j^*}$ we must have that

$$\bar{e}_{j^*}(v) - e_{j^*}(v) \leq \bar{e}_k(v) - e_k(v) \qquad (\forall k \neq j^*)\,, \tag{20}$$

where if $c' = c$ then $1 \leq k \leq c'$, while if $c' < c$ then $0 \leq k \leq c'$. (Note that Eq. (20) holds since otherwise v would be placed in $\widetilde{V}_k$.) Eq. (20) will be useful when we consider the following subcases (which refer to $v \in \widetilde{V}_{j^*}$).

- We first consider the subcase in which $j^* = 0$ (which may occur only when $c' < c$). In this subcase, since $\bar{e}_{j^*}(v) - e_{j^*}(v) = 0 - 0 = 0$, for every $k \in [c']$ we have that $\bar{e}_k(v) \geq e_k(v)$. On the other hand, since $x(v) = \sum_{k=1}^{c'} e_k(v) \geq N/2^{i^*}$, there exists at least one index $k^* \in [c']$ such that $e_{k^*}(v) \geq N/(c2^{i^*})$. Since $\bar{e}_{k^*}(v) \geq e_{k^*}(v)$, we have that $\bar{e}_{k^*}(v) \geq N/(c2^{i^*})$ as well. This also implies that $|V_{k^*}|/N \geq (c2^{i^*})^{-1}$, and since we assume that $|C_{k^*}|/|S| \geq \frac{1}{2}|V_{k^*}|/N$, we have that $|C_{k^*}|/|S| \geq (2c2^{i^*})^{-1}$. Recall that $|T^2_{i^*}| = \Theta(2^{i^*})$, and that $|S_{i^*}| = \Theta(\min\{2^{i^*}, 1/(\epsilon 2^{i^*})\}) = \Theta(2^{i^*})$, since $i^* \leq \ell/2+2$ (where $\ell = \log(1/\epsilon)$).

  Now, if $|C_{k^*} \cap \Gamma(v)| \geq |C_{k^*}|/2$, then, with high probability, the sample S_{i^*} contains a vertex w in $C_{k^*} \cap \Gamma(v)$ (since $|C_{k^*}| = \Omega(|S|/2^{i^*})$), and $T^2_{i^*}$ contains a vertex u in $V_{k^*} \setminus \Gamma(v)$ (since $\bar{e}_{k^*}(v) = \Omega(N/2^{i^*})$). Otherwise (i.e., $|C_{k^*} \setminus \Gamma(v)| \geq |C_{k^*}|/2$), with high probability, S_{i^*} contains a vertex w in $C_{k^*} \setminus \Gamma(v)$, and $T^2_{i^*}$ contains a vertex u in $V_{k^*} \cap \Gamma(v)$ (since $e_{k^*}(v) = \Omega(N/2^{i^*})$). In either cases, $w \in C_{k^*}$ and $u \in V_{k^*}$, which implies $(u,w) \in E$, and $w \in \Gamma(v)$ iff $u \notin \Gamma(v)$, which implies that $|\{(u,w),(w,v),(u,v)\} \cap E| = 2$.

  In the subsequent subcases we assume that $j^* > 0$.

- We next consider the subcase in which both $\bar{e}_{j^*}(v) \geq N/2^{i^*+1}$ and $e_{j^*}(v) \geq N/2^{i^*+2}$ hold. Setting $k^* \leftarrow j^*$, we reach a situation as in the first subcase (since $\bar{e}_{k^*}(v) = \Omega(N/2^{i^*})$ and $e_{k^*}(v) = \Omega(N/2^{i^*})$), and we are done as in the first subcase (while noting that first subcase does not rely on $j^* \neq k^*$).

- The next subcase refers to $\bar{e}_{j^*}(v) \geq N/2^{i^*+1}$ and $e_{j^*}(v) < N/2^{i^*+2}$. In this subcase $\bar{e}_{j^*}(v) - e_{j^*}(v) > 0$ and so it can occur only when $c' = c$ (since otherwise v would be placed in $\widetilde{V}_0$, whereas here $j^* \neq 0$)). The fact that $\bar{e}_{j^*}(v) - e_{j^*}(v) \geq N/2^{i+2}$ implies that, for every

$k \in [c'] \setminus \{j^*\}$, it holds that $\bar{e}_k(v) \geq e_k(v) + \bar{e}_{j^*}(v) - e_{j^*}(v) \geq N/2^{i^*+2}$. Similarly to the previous subcase, we know that $|C_k|/|S| \geq 1/2^{i^*+3}$ for all k, and we have that $|S_{i^*}| = \Theta(2^{i^*})$ (as well as $|T^2_{i^*}| = \Theta(2^{i^*})$).

If there exists $k^* \in [c']$ such that $|C_{k^*} \cap \Gamma(v)| \geq |C_{k^*}|/2$, then with high probability, S_{i^*} contains a vertex in $C_{k^*} \cap \Gamma(v)$, and $T^2_{i^*}$ contains a vertex in $V_{k^*} \setminus \Gamma(v)$. Otherwise (i.e., $|C_k \setminus \Gamma(v)| \geq |C_k|/2$ for every $k \in [c']$), with high probability, for every $k \in [c']$, the sample S_{i^*} contains a vertex in $C_k \setminus \Gamma(v)$, and recalling that $c' = c$ we obtain evidence (in the form of an independent set of size $c+1$) that G is not in $\mathcal{CC}^{\leq c}$.

- Lastly, we consider the subcase in which $\bar{e}_{j^*}(v) \leq N/2^{i^*+1}$. Since $\bar{e}_{j^*}(v) + \sum_{k \in [c'] \setminus \{j^*\}} e_k(v) = x(v) > N/2^{i^*}$, we obtain $\sum_{k \in [c'] \setminus \{j^*\}} e_k(v) \geq N/2^{i^*+1}$. In such a case, there exists a $k^* \in [c'] \setminus \{j^*\}$ for which $e_{k^*}(v) \geq N/(c2^{i^*+1})$. If $e_{j^*}(v) \geq N/(c2^{i^*+2})$, then with high probability, $T^2_{i^*}$ contains one vertex u in $V_{k^*} \cap \Gamma(v)$ and one vertex u' in $V_{j^*} \cap \Gamma(v)$, while S_{i^*} contains one vertex w in C_{k^*} and one vertex w' in C_{j^*}, and we have evidence that G is not a union of cliques (since $(v,u), (v,u'), (u,w), (u',w') \in E$ whereas $(w,w') \notin E$, and all five vertex pairs are inspected by the algorithm).[20] Otherwise (i.e., $e_{j^*}(v) < N/(c2^{i^*+2})$), by Eq. (20), we have that $\bar{e}_{k^*}(v) \geq e_{k^*}(v) + \bar{e}_{j^*}(v) - e_{j^*}(v) \geq N/(c2^{i^*+2})$, and we are in essentially the same situation as the first subcase (since we have $e_{k^*}(v) = \Omega(N/2^{i^*})$ and $\bar{e}_{k^*}(v) = \Omega(N/2^{i^*})$).

It remains to deal with the assumption that $|C_j|/|S| \geq \frac{1}{2}|V_j|/N$ holds for every j that satisfies $|V_j| \geq \frac{\epsilon^{1/2}}{2c}N$. To this end, we add one more phase in the choice of S (where we think of this phase as taking place before phase $c+1$ that was used in the foregoing discussion to bound $|R|$). Let S' denote the vertices selected in the first c phases and let S'' be the vertices selected in the additional phase, where $|S''| = 4|S'|$. Let $C'_1, \ldots, C'_{c'}$ be the cliques in the subgraph induced by S', and for each $1 \leq j \leq c'$ let V'_j be the vertices that neighbor all vertices in C'_j and no other vertices in S'. In the sample S'', let $C''_j = S'' \cap V'_j$. By a multiplicative Chernoff bound, with high probability over the choice of S'', it holds that $|C''_j|/|S''| \geq (3/4)|V'_j|/N$ for every j that satisfies $|V'_j| \geq \frac{\epsilon^{1/2}}{2c}N$. Assume that this is, in fact, the case. Then, we define $C_j = C'_j \cup C''_j$ and $V_j = \{v : \Gamma(v) \cap (S' \cup S'') = C_j\}$.

If there is any new clique in S'' then it corresponds to a small set of vertices (since the set of vertices that do not belong to any V'_j is small).[21] Using the fact that S is the union of S', S'' and the sample selected in phase $c+1$, we have $|S| < (3/2)|S''|$ (since $|S''| = 4|S'|$ and $|S'| = c \cdot (|S| - |S'| - |S''|)$) and $|C_j|/|S| \geq (3/4)|C''_j|/|S''| \geq (3/4) \cdot (3/4)|V'_j|/N$. Using $V_j \subseteq V'_j$, we get that $|C_j|/|S| > \frac{1}{2}|V_j|/N$ for every $|V_j| \geq \frac{\epsilon^{1/2}}{2c}N$. ■

[20] Actually, note that it also holds that $(u',w) \notin E$, and thus we obtain evidence in the form of the four vertex pairs $(v,u), (v,u'), (u,w), (u',w)$. Note that we can obtain evidence in the form of three vertex pairs by considering either $(v,u), (u',w), (v,w)$ or $(v,u), (u,w), (v,w)$.

[21] Indeed, the sizes of the sets V'_j behave as the sizes of the sets V_j, which were analyzed in the beginning of this proof. Also note that this additional clique may causes the algorithm to reject (whenever it causes the total number of cliques to exceed c).

References

[A] N. Alon. On the number of subgraphs of prescribed type of graphs with a given number of edges. *Israel J. Math. 38*, pages 116–130, 1981.

[AFKS] N. Alon, E. Fischer, M. Krivelevich and M. Szegedy. Efficient Testing of Large Graphs. *Combinatorica*, Vol. 20, pages 451–476, 2000.

[AFN] N. Alon, E. Fischer, and I. Newman. Testing of bipartite graph properties. *SIAM Journal on Computing*, Vol. 37, pages 959–976, 2007.

[AFNS] N. Alon, E. Fischer, I. Newman, and A. Shapira. A Combinatorial Characterization of the Testable Graph Properties: It's All About Regularity. In *38th STOC*, pages 251–260, 2006.

[AS] N. Alon and A. Shapira. A Characterization of Easily Testable Induced Subgraphs. In *15th SODA*, pages 935–944, 2004.

[BHR] E. Ben-Sasson, P. Harsha, and S. Raskhodnikova. 3CNF properties are hard to test. *SIAM Journal on Computing*, Vol. 35 (1), pages 1–21, 2005.

[BT] A. Bogdanov and L. Trevisan. Lower Bounds for Testing Bipartiteness in Dense Graphs. In *IEEE Conference on Computational Complexity*, pages 75–81, 2004.

[CEG] R. Canetti, G. Even and O. Goldreich. Lower Bounds for Sampling Algorithms for Estimating the Average. *IPL*, Vol. 53, pages 17–25, 1995.

[F01] E. Fischer. The art of uninformed decisions: A primer to property testing. *Bulletin of the European Association for Theoretical Computer Science*, Vol. 75, pages 97–126, 2001.

[F04] E. Fischer. On the strength of comparisons in property testing. *Inform. and Comput.*, Vol. 189 (1), pages 107-116, 2004.

[GGR] O. Goldreich, S. Goldwasser, and D. Ron. Property testing and its connection to learning and approximation. *Journal of the ACM*, pages 653–750, July 1998.

[GR02] O. Goldreich and D. Ron. Property Testing in Bounded Degree Graphs. *Algorithmica*, Vol. 32 (2), pages 302–343, 2002.

[GR08] O. Goldreich and D. Ron. On Proximity Oblivious Testing. *ECCC*, TR08-041, 2008.

[GT] O. Goldreich and L. Trevisan. Three theorems regarding testing graph properties. *Random Structures and Algorithms*, Vol. 23 (1), pages 23–57, August 2003.

[GnRn] M. Gonen and D. Ron. On the Benefit of Adaptivity in Property Testing of Dense Graphs. In *Proc. of RANDOM'07*, LNCS Vol. 4627, pages 525–539, 2007.

[R01] D. Ron. Property testing. In *Handbook on Randomization, Volume II*, pages 597–649, 2001. (Editors: S. Rajasekaran, P.M. Pardalos, J.H. Reif and J.D.P. Rolim.)

[RaSm] S. Raskhodnikova and A. Smith. A note on adaptivity in testing properties of bounded-degree graphs. *ECCC*, TR06-089, 2006.

[RuSu] R. Rubinfeld and M. Sudan. Robust characterization of polynomials with applications to program testing. *SIAM Journal on Computing*, 25(2), pages 252–271, 1996.