

# Implementing Huge Sparse Random Graphs

Moni Naor\* and Asaf Nussboim†

Department of Computer Science and Applied Mathematics  
Weizmann Institute of Science, Rehovot, Israel.  
{moni.naor, asaf.nussbaum}@weizmann.ac.il.

August 29, 2007

## Abstract

Consider a scenario where one desires to simulate the execution of some graph algorithm on random input graphs of huge, perhaps even exponential size. Sampling and storing these huge random graphs is clearly infeasible, but can they be emulated by ‘random looking’ graphs that are efficiently computable? Recently, Goldreich et al. [8], and Naor et al. [13] presented efficient implementations of the canonical (dense) random graphs  $G(N, p)$  where  $N = 2^n$  labeled vertices are fixed and each edge independently appears with probability  $p = p_n$ .

We continue this line of research by emulating *sparse*  $G(N, p)$  graphs. The reasonable model for accessing the latter is by efficiently evaluating the entire (small) neighborhood  $\Gamma(v)$  in response to a query-vertex  $v$ . We cover a wide range of densities including random graphs’ famous threshold density for containing a giant component ( $p \sim 1/N$ ), and for achieving connectivity ( $p' \sim \ln N/N$ ). Our graphs faithfully emulate random graphs in the sense that they are indistinguishable from  $G(N, p)$  graphs from the view of any efficient algorithm that inspects the graph by neighborhood queries of its choice. The main challenges we meet are (i) efficiently approximating the degree sequence of random graphs, and (ii) retrieving the neighborhood  $\Gamma(v)$ , without sequentially deciding the adjacency of  $v$  w.r.t. each of the exponentially many other vertices.

---

\*Partly supported by a grant from the Israel Science Foundation.

†Partly supported by the Minerva Foundation 2-8495.

# 1 Introduction

Consider a scenario where one desires to simulate the execution of some graph algorithm on random input graphs of huge size, perhaps even exponentially large in the input length,  $n$ , of the corresponding algorithms. Sampling and storing these huge random graphs is clearly infeasible, but can they be emulated by ‘random looking’ graphs that are efficiently computable?

This question of emulating huge random graphs continues a rich body of research regarding the implementation of huge random objects. Huge random objects can often be faithfully replaced by some ‘random-looking’ counterparts that are sampled from distributions of significantly smaller support, and can thus be efficiently utilized (namely, using polynomially bounded resources). In general, huge objects are not represented explicitly, but rather by an efficient procedure that evaluates queries regarding the object (e.g. input-output queries on functions). These queries are evaluated using a succinct ( $poly(n)$ -length) representation of the object called seed. Thus, random looking distributions are sampled by randomly picking the seed.

Examples of highly influential random looking objects, which, in fact, underly the foundations of modern cryptography, are pseudorandom functions  $f : \{0, 1\}^n \rightarrow \{0, 1\}^n$  and pseudorandom permutations over similar domains. The former were defined and constructed by Goldreich, Goldwasser and Micali [7] (under the necessary assumption that one-way functions exist<sup>1</sup>), and the latter were provided by Luby and Rackoff [11] (based on [7]). The criterion introduced in [7] for faithful emulation of random objects is computational indistinguishability. Namely, no efficient distinguishing algorithm that inspects the function (permutation, resp.) via a sequence of input-output queries of its choice, can tell with probability significantly better than  $\frac{1}{2}$  whether the function (permutation, resp.) is sampled from the pseudorandom distribution or from the uniform distribution over functions (permutations, resp.)  $f : \{0, 1\}^n \rightarrow \{0, 1\}^n$ .

## 1.1 Implementing Huge Random Graphs

Recently, Goldreich et al. [8], and Naor et al. [13] studied the implementation of huge random graphs. They considered the canonical random graphs  $G(N, p_n)$ , where  $N = 2^n$  labeled vertices are fixed and each edge appears with probability  $p = p_n$  independently of all other edges. These works focused on relatively dense graphs, where the natural access model is via edge-queries that inquire whether a specific edge appears in the graph.

In contrast, the focus of this work is sparse graphs. The latter term refers to a wide range of densities  $p$ , including the Erdős-Rényi threshold density for containing a giant component ( $p \sim 1/N$  [5]), and for achieving connectivity ( $p' \sim \ln N/N$  [4]). As edge-queries rarely detect any adjacency in sparse graphs, the reasonable model for accessing them is by providing the entire neighborhood (namely, the entire list of adjacent vertices)  $\Gamma(v)$  in response to a query-vertex

---

<sup>1</sup>A one-way function is an efficiently computable function  $h : \{0, 1\}^* \rightarrow \{0, 1\}^*$  that cannot be inverted efficiently on random inputs with non-negligible probability of success.

$v$  (this neighborhood-queries model is the common one in the context of sparse graphs, and in particular, in the field of testing graph properties without inspecting the entire graph [9]).

Supporting neighborhood-queries is far from trivial, as one has to specify which of the *exponentially* many potential vertices indeed appears in  $\Gamma(v)$ . In addition, since the graphs are undirected, consistency is required in the sense that  $u \in \Gamma(w)$  iff  $w \in \Gamma(u)$ . In particular, previous works [8, 13] implement a graph via a Boolean function  $f$ , where  $f(e) = 1$  iff the edge  $e$  appears in the graph. Using pseudorandom Boolean functions [7] guarantees efficiency (in the usage of randomness and memory), but eventually fails in our context, as supporting even a single neighborhood-query requires the evaluation of  $f$  on exponentially many inputs.

To overcome this, a different approach was applied by Goldreich et al. to support neighborhood-queries [8]. They construct sparse  $d$ -regular graphs that achieve computational pseudo-randomness w.r.t. the uniform distribution over  $d$ -regular graphs. They rely on the fact that the views observed by efficient distinguishers (that examine sparse random regular graphs) are typically cycle-free and are symmetric (w.r.t. the role of different vertices in the view). Thus, randomly shuffling the vertex-names of any single large-girth  $d$ -regular graph will produce a random looking cycle-free view, as desired. Cycle-free views characterize *non-regular* sparse random graphs too. Yet our main challenge in the present work is (not only to construct graphs that produce cycle-free views, but) mainly to properly approximate the degree distribution of  $G(N, p)$ .

## 1.2 Our Contribution

We construct computationally pseudorandom graphs w.r.t. neighborhood-queries, under the necessary assumption that one-way functions (OWF) exist. Pseudorandomness is achieved (even) w.r.t. adaptive distinguishers, that may choose the next query depending on previous replies (in particular, the next query vertex may appear in a previous reply). The graphs produced by our construction are always simple (no self-loops or multi-edges) and undirected (so  $u \in \Gamma(w)$  iff  $w \in \Gamma(u)$  holds for any vertex pair  $u, v$ ). Our results hold for the entire range of densities where typical degrees are *poly*( $n$ )-bounded, namely, for arbitrary  $p_n \leq \frac{n^{O(1)}}{N}$ . These results can be easily generalized from order  $N = 2^n$  into any super-polynomial order  $n^{\omega(1)} \leq N \leq 2^{\text{poly}(n)}$ .

Standard pseudorandomness arguments imply the necessity of the OWF assumption whenever  $p \geq \frac{1}{Nn^{O(1)}}$ . Thus, the OWF assumption is necessary to capture the threshold densities for containing a giant component ( $p \sim 1/N$ ), and for achieving connectivity ( $p' \sim \ln N/N$ ). For smaller densities  $p \leq \frac{1}{Nn^{\omega(1)}}$ , OWFs are no longer needed since  $G(N, p)$ -views rarely include any adjacencies, and therefore the empty graph provides the desired pseudorandom implementation.

We remark that one may consider a weaker notion of efficiency, under which it is possible to handle (again, under the OWF assumption) higher densities:

here neighborhoods are allowed to have super-polynomial size, yet each query is still handled in polynomial time in the size of the reply. In this case, pseudo-randomness is achieved against polynomially bounded distinguishers that may query the degree of a given vertex  $v$  and may receive a random neighbor of  $v$ . With respect to this weaker efficiency definition, our construction applies whenever the density is negligible. The latter roughly captures the entire range where the edge-queries model is no longer reasonable (so neighborhood-queries are used instead). We stress that subsequent definitions and results in this paper relate only to the stronger (and more standard) notion of efficiency.

### 1.2.1 Description of our Construction.

We first provide a costly interim construction that emulates  $G(N, p)$  well, and then ‘de-randomize’ our interim implementation to obtain an efficient one. In the interim construction,  $G_{\text{IBin}}$ , the degree of each specific vertex has Binomial distribution  $\text{BIN}(N-1, p)$  (just as in  $G(N, p)$ ). However, unlike the  $G(N, p)$  case, all the degrees in  $G_{\text{IBin}}$  are independent of each other<sup>2</sup> (for instance, the sum of degrees in  $G_{\text{IBin}}$  is allowed to be odd). Given the degrees, edges are assigned to the graph via the traditional configuration method (Bollobás [2]) where each vertex of degree  $d$  is associated with  $d$  unique ‘ports’. A uniformly random matching over the ports decides the edges of the graph s.t. two vertices are adjacent iff any of their ports are matched (self-loops and multi-edges are ignored<sup>3</sup>).

The indistinguishability of  $G_{\text{IBin}}$  from  $G(N, p)$  is established by showing that the distribution of  $G_{\text{IBin}}$ -replies is statistically close<sup>4</sup> to the corresponding distribution in the  $G(N, p)$  case - as long as the number of queries is  $\text{poly}(n)$ -bounded. To this end, it is observed that in the  $G(N, p)$  case the size of the next reply  $\Gamma_j$  has Binomial distribution, with each specific vertex being equally likely to appear in  $\Gamma_j$  (this holds regardless of the previous replies  $\Gamma_1, \dots, \Gamma_{j-1}$ ). Thus, the main technical part of the proof is to analyze the distribution of the next reply in the  $G_{\text{IBin}}$  model and to establish its closeness (up to a negligible difference) to the  $G(N, p)$  case.

The interim construction is ‘de-randomized’ as follows. Neighborhood-queries are handled by using random-looking functions that efficiently support interval-queries where on interval-query  $(\alpha, \beta)$  the entire sum  $\sum_{\alpha \leq x \leq \beta} f(x)$  is retrieved. Implementing such functions is due to Goldreich et al. [8] and to Naor and Reingold (in [6]). Specifically, we use a Boolean function  $f$  over  $N(N-1)$  inputs that are partitioned into  $N$  blocks of size  $N-1$ . The sum of  $f$  over the  $v$ ’th block corresponds to the number of ports  $d_v$  of vertex  $v$ . Thus, the integers  $1, \dots, d_1$  are the ports of the first vertex,  $d_1 + 1, \dots, d_1 + d_2$  are the ports of the second

<sup>2</sup>Thus the notation IBin stands for the fact that the joint distribution of the degrees  $(d_1, \dots, d_n)$  is the product-distribution of  $N$  Binomial distributions.

<sup>3</sup>When the total number of ports is odd, a single port remains unmatched and thus induces no edge.

<sup>4</sup>The statistical distance between two distributions  $\mathcal{D}_1, \mathcal{D}_2$  is defined as  $\frac{1}{2} \sum_x |\Pr_{\mathcal{D}_1}[x] - \Pr_{\mathcal{D}_2}[x]|$  where the sum ranges over the entire sample space.

vertex, etc.. To retrieve the neighborhood  $\Gamma(v)$ , we use interval-queries over  $f$  to identify the exact set of ports that  $v$  possesses, and then apply the random matching on these ports. To resemble  $G_{\text{IBin}}$ , the aforementioned construction [8, 6] guarantees our functions  $f$  to be computationally indistinguishable (w.r.t. interval-queries) from the (truly random) functions that correspond to  $N(N - 1)$  independent Bernoulli trials with success probability  $p$ .

Finally, as the total number of ports is typically exponentially large, the required random matching over the ports is formalized as a random involution (with no fixed-points) and Naor and Reingold’s construction of pseudorandom involutions [14] is applied to efficiently match the ports and decide  $\Gamma(v)$ .

### 1.2.2 Achieving Almost $k$ -Wise Independence.

We briefly discuss  $(k, \epsilon)$ -wise independence, which is an alternative (and incomparable) criterion for being a good ‘random looking’ graph. Meeting this criterion means that for any  $k$  vertices  $u_1, \dots, u_k$  (fixed in advance) the distribution of the neighborhoods  $\Gamma(u_1), \dots, \Gamma(u_k)$  is within  $\epsilon$  statistical distance from the corresponding distribution in the  $G(N, p)$  case.

Our construction can achieve  $(k, \epsilon)$ -wise independence for any prescribed  $\text{poly}(n)$  bounded value of  $k$  and any prescribed exponentially small  $\epsilon$ . This is done by slightly modifying the implementation of the pseudorandom involutions and the implementation of the pseudorandom functions that support interval-queries. Rather than using computationally pseudorandom bits for the two latter implementations (as originally done in [6, 8, 14]),  $k'$ -wise independent bits are used instead. The latter refer to distributions s.t. for any fixed  $k'$  bits, their joint distribution is precisely uniform. It can be shown that taking some  $k' \leq \text{poly}(n, k)$  suffices for our resulting graphs to be  $(k, \epsilon)$ -wise independent. Thus by applying known efficient constructions of  $k'$ -wise independent bits (cf. [1] chp. 15) the efficiency of our modified construction is retained. Constructions of bits that are both pseudorandom and  $k'$ -wise independent are easily given (by combining the original constructions for each type), thus providing graphs which are simultaneously pseudorandom as well as  $(k, \epsilon)$ -wise independent.

### 1.2.3 Emulating Related Models of Random Graphs.

Consider the original Erdős-Rényi models  $G(N, M)$  and  $\{G(N, t)\}_t$  which are closely related to  $G(N, p)$  graphs. In both models  $N$  labeled vertices are fixed (as before), with  $G(N, M)$  being the uniform distribution over all graphs with precisely  $M$  edges, whereas  $\{G(N, t)\}_t$  is the random graph process, where the initial graph  $G(N, 0)$  is empty and at each time step,  $G(N, t + 1)$  is obtained from  $G(N, t)$  by adding a uniformly random edge (that hasn’t been chosen before). Thus,  $\{G(N, t)\}_t$  at time  $t = M$  is identical to  $G(N, M)$ , and it is well known that the combinatorial properties of  $G(N, p)$  and  $G(N, M)$  graphs are very similar when  $p \sim M/\binom{N}{2}$ .

We demonstrate how to emulate  $G(N, M)$  and  $\{G(N, t)\}_t$  graphs where for  $\{G(N, t)\}_t$  graphs the reasonable types of queries are: i) which edge was added

at time step  $t$ , and ii) whether some specific edge already appears at time  $t$ . The efficiency and pseudorandomness of the following constructions is easy to establish given the main theorems of this paper.

The first construction (appropriate for dense graphs too) uses a pseudorandom bijection  $\sigma$  from the set of all possible time-steps  $\{1, \dots, \binom{N}{2}\}$  to the set of all  $\binom{N}{2}$  potential edges. Thus,  $\sigma(t)$  is the edge joined to the graph at time  $t$ , and the edge  $e$  appears in the graph at time  $t$  iff  $\sigma^{-1}(e) \leq t$ . Similarly,  $G(N, M)$  is emulated by including in the graph precisely those edges s.t.  $\sigma^{-1}(e) \leq M$ .

For sparse graphs, the  $\{G(N, t)\}_t$  process is bounded a-priori at some  $T = n^{\Theta(1)}N$  time, and neighborhood-queries should be supported. To this end, the main construction of this paper is used with the following two adaptations. i) We first (trivially) modify the construction of the pseudorandom range-summable functions, s.t. precisely  $2T$  ports are produced (instead of a Binomially distributed number of ports). Deciding the edge-set of the resulting graph, and supporting neighborhood-queries in a pseudorandom manner is done as before. ii) In addition, we use a pseudorandom bijection  $\sigma$  from the set of all possible time-steps  $\{1, \dots, T\}$  to the set of  $T$  edges that match the ports in our construction s.t.  $\sigma(t)$  is the edge joined to the graph at time  $t$ . Deciding whether the edge  $\{u, v\}$  already appears at time  $t$ , is done by enumerating all  $\text{poly}(n)$  ports  $\rho_i$  of  $u$ , and for each of them checking whether  $\rho_i$  was matched to a port of  $v$  prior to time  $t$ . Unfortunately, this account of time steps fails to ignore double-edges and self loops, in contrast with  $\{G(N, t)\}_t$ , (but in line with the variant of  $\{G(N, t)\}_t$  where at each step a uniformly random vertex-pair is added, with repetitions allowed).

## 2 Preliminaries

This section provides definitions and notations. The main definitions are of the models  $G_{\text{IBin}}$  and  $G_{\text{IBin}}^{\text{OTF}}$  (under ‘graphs and configurations’), and the definitions (derived from [8]) of pseudorandomness not only for graphs, but also for interval-summable functions and for involutions.

### 2.1 Basics, Arithmetics and Asymptotic Analysis

Efficient procedures are algorithms that run in worst-case polynomial time in their input length  $n$ . Throughout, all graphs have  $N = 2^n$  vertices. Negligible terms  $\epsilon(n)$  are ones that vanish faster than the reciprocal of any polynomial (for all  $j$ ,  $|\epsilon(n)| = o(n^{-j})$ ). The notation  $X(1 \pm \delta)$  stands for some term  $E$  s.t.  $X(1 - \delta) \leq E \leq X(1 + \delta)$ , and the notation  $A \sim B$  implies that  $A = B(1 \pm \epsilon)$  for some negligible  $\epsilon$ . We often use the fact that  $\sim$  is a transitive relation that behaves well under summation, multiplication and division. Some of the inequalities used throughout hold only for sufficiently large (yet very reasonable) values of  $n$ . We use the notation  $[m]$  for  $\{1, \dots, m\}$ .

## 2.2 Graphs and Configurations

We consider only simple, undirected graphs (no self-loops or multi-edges allowed), over the vertex-set  $\{1, \dots, N\}$ . A port sequence is any sequence  $\vec{t} = (t_1, \dots, t_N) \in \{0, \dots, N-1\}^N$ , regardless of whether  $\vec{t}$  is indeed a graphic-sequence (namely the degree sequence of some  $N$ -vertex graph). In particular we allow the sum  $\sum_{v=1}^N t_v$  to be odd. The term ‘degree’ and the notation ‘ $d_v$ ’ are abused to refer not only to degrees of vertices, but also to the number of ports that a vertex  $v$  has in the configurational model (to avoid confusion, this notation is never used in sections that concurrently discuss degree sequences and port sequences).

**The  $G_{\text{IBin}}$  model.** In this model each vertex  $v$  is associated with  $d_v$  unique ports where the number of ports has Binomial distribution  $d_v \sim \text{BIN}(N-1, p)$ , and all the random variables  $d_v$  are independent of each other. If the sum  $\sum_{v=1}^N d_v$  is odd, a single ‘parity-vertex’ with a single port is added (to obtain an even sum). A matching over the ports is later chosen uniformly at random (among all possible matchings), s.t. two vertices are adjacent iff any of their ports are matched. Self-loops, multi-edges, and (possibly) the edge connected to the parity vertex are all ignored in the final graph.

**The  $G_{\text{IBin}}^{\text{OTF}}$  model.** Here the port sequence is produced as in  $G_{\text{IBin}}$ , but the matching over the ports is decided ‘on-the-fly’, during a single interaction with some distinguishing algorithm  $C$ . On query-vertex  $v$ , the available ports associated with  $v$  are matched one after the other to a uniformly random available port. After the execution of  $C$  terminates, the remaining available ports are uniformly matched at random. It is not too hard to see, that the graph distribution produced by the two models is identical.

## 2.3 Computational Pseudorandomness

As in [8], the following definitions capture pseudorandomness not only for graphs (w.r.t. neighborhood-queries) but also for interval-summable functions and for involutions. To this end, we use query representation functions.

**Query representation functions (QRFs).** Given the type of queries we wish to support, we represent each specific object  $o$  (e.g. a single graph) by a specific function  $f_o$  s.t. evaluating  $f_o$  on a single input corresponds to supporting a single (possibly complex) query over  $o$ . We call  $f_o$  the QRF of  $o$ . For instance to support neighborhood queries over  $N$  vertex graphs, each specific graph  $g$  is represented by a function  $f_g : [N] \rightarrow 2^{[N]}$  s.t. for any vertex  $v$ ,  $f_g(v) = \Gamma(v)$ . Similarly, to support interval-queries over Boolean functions with domain  $[M]$ , each specific function  $h$  is represented by another function  $f'_h : [M] \times [M] \rightarrow [M]$  s.t.  $f'_h(\alpha, \beta) = \sum_{\alpha \leq x \leq \beta} h(x)$  for all  $\alpha, \beta \in [M]$ . The QRF for involutions (w.r.t. input-output queries) is, of course, the original involution itself.

**Pseudorandomness w.r.t. complex queries.** Note that any distribution  $\mathcal{D}$  over the original objects induces a corresponding distribution  $\mathcal{F}_{\mathcal{D}}$  over the QRFs. Consequently, pseudorandomness of the original objects  $\mathcal{D}$  w.r.t. a given type of

queries, reduces to the pseudorandomness of the QRFs  $\mathcal{F}_{\mathcal{D}}$  w.r.t. simple input-output queries. We thus define pseudorandomness w.r.t. complex queries simply as pseudorandomness (in the classical sense of GGM [7]) of the corresponding QRFs. For a given sequence of densities  $\{p_n\}_{n \in \mathbb{N}}$   $p_n \in [0, 1]$  the following definitions are used:

- *Neighborhood queries pseudorandom graphs.* Let  $\{\mathcal{G}_n\}_{n \in \mathbb{N}}$  be a sequence of distributions, where each  $\mathcal{G}_n$  is taken over  $N$ -vertex graphs. Then  $\{\mathcal{G}_n\}_{n \in \mathbb{N}}$  is neighborhood-queries pseudorandom w.r.t.  $\{G(N, p_n)\}_{n \in \mathbb{N}}$  if the neighborhood QRFs  $\{\mathcal{F}_{\mathcal{G}_n}\}_{n \in \mathbb{N}}$  induced by  $\{\mathcal{G}_n\}_{n \in \mathbb{N}}$  is pseudorandom w.r.t. the neighborhood QRFs induced by  $\{G(N, p_n)\}_{n \in \mathbb{N}}$ .

- *Interval-sum queries pseudorandom functions.* Consider a sequence of distributions  $\{\mathcal{H}_n\}_{n \in \mathbb{N}}$  where each  $\mathcal{H}_n$  is taken over Boolean functions with domain  $D_n$ . Then  $\{\mathcal{H}_n\}_{n \in \mathbb{N}}$  is interval-sum query pseudorandom if its interval-sum QRFs are pseudorandom w.r.t. the interval-sum QRFs of the truly random functions. The latter refer to the distribution over Boolean functions  $h_n$  with domain  $D_n$ , where for each input  $x$  we have  $h_n(x) = 1$  with probability  $p_n$  independently of the value of  $h_n$  over other inputs.

- *Pseudorandom involutions.* A sequence of integers  $\{M_n\}_{n \in \mathbb{N}}$  forms proper involutions domains if all  $M_n$  are even and  $M_n = n^{\omega(1)}$ .<sup>5</sup> A sequence of distributions  $\{\Pi_n\}_{n \in \mathbb{N}}$  over involutions with no fixed-points  $\pi_n : [M_n] \rightarrow [M_n]$  are pseudorandom involutions if it is pseudo-random w.r.t. the sequence of uniform distributions over all involutions with no fixed-points and same domains  $[M_n]$ .

Finally, recall that efficiently constructing pseudorandom functions is possible iff one-way functions (OWFs) exist, which are efficiently computable functions  $f : \{0, 1\}^* \rightarrow \{0, 1\}^*$  that cannot be inverted efficiently on uniformly random inputs with non-negligible probability of success.

### 3 Our Main Construction

This section formally describes our construction of sparse pseudorandom graphs.

We first define the range of ‘proper densities’  $p$  that our arguments handle. To ensure that all the degrees are  $poly(n)$ -bounded, proper densities are upper bounded. Our techniques also require a lower bound which guarantees that the total number of ports is (almost surely) (i) super-polynomial (in  $n$ ) and (ii) extremely close to its expectation. These facts are (i) frequently used while proving the similarity of the models  $G_{\text{IBin}}$  and  $G(N, p)$ , and (ii) used to validate the usage of the pseudorandom involutions. For densities too small to be proper,  $G(N, p)$ -views rarely include any adjacencies, so  $G(N, p)$  graphs are emulated well by the empty graph.

**Definition 1 (Proper density)** *A sequence of densities  $\{p_n\}_{n \in \mathbb{N}}$  is proper if for all  $n$ ,  $0 < p_n < 1$  and  $\frac{(\lg N)^{\omega(1)}}{N^2} \leq p_n \leq \frac{(\lg N)^{O(1)}}{N}$ .*

<sup>5</sup>A super-polynomial domain is necessary to ensure pseudorandomness w.r.t. polynomial adversaries.

We next present the main constructions we use as sub-routines.

**Theorem 1 ([8, 6] Interval-summable functions)** *Let  $\{p_n\}_{n \in \mathbb{N}}$  be proper. Assuming the existence of one-way functions, then [8, 6] provide interval-summable queries pseudorandom functions with domains  $D_n = [N] \times [N - 1]$ .*

**Theorem 2 ([14] Pseudorandom involutions)** *Let  $\{M_n\}_{n \in \mathbb{N}}$  be a proper involutions domains. Assuming the existence of one-way functions, [14] provide pseudorandom involutions w.r.t. the domains  $\{[M_n]\}_{n \in \mathbb{N}}$ .*

We remark that the original [14] construction handles only domains of size which is a power of 2. The adaptation to the more general case is discussed in Appendix 2, section 6.3.

Our main construction is given bellow. The underlying intuition is discussed in the introduction.

**Construction 1 (Implementing sparse random graphs)** *On input  $(1^n, p)$ , construct a graph on vertex-set  $[N]$  as follows. If  $p \leq N^{-1.5}$  pick the empty graph so each query is replied with the empty set. Otherwise,*

- **Sampling** - *Pick an Interval-summable function  $f$  over domain  $X = [N(N - 1)]$  with parameter  $p$ , as in Theorem 1. Set  $E = \sum_{x \in X} f(x)$  to be the total number of ports. If  $E$  is odd increase it by 1 (adding the parity port). Sample a pseudorandom involution (with no fixed points)  $\pi : [E] \rightarrow [E]$ , as in Theorem 2.*
- **Supporting neighborhood-queries** - *On query-vertex  $v$ :*
  - (i) *Compute  $S_v = \sum_{x=1}^{(v-1)(N-1)} f(x)$ ,  $S'_v = \sum_{x=1}^{v(N-1)} f(x)$  and  $d_v = S'_v - S_v$  (thus  $S_v + 1, \dots, S_v + d_v$  are the ports associated with  $v$ ).*
  - (ii) *For  $i = 1, \dots, d_v$  compute  $T_i = \pi(S_v + i)$  ( $T_i$  is the port matched with  $S_i$ ). Unless  $T_i$  is the parity-port, conduct a binary search to decide to which vertex  $u_i$  the port  $T_i$  belongs (the space of the search is the vertex-set  $[N]$ . At each stage of the search step (i) is invoked to check whether  $T_i$  belongs to  $u$  or to a previous or consequent vertex  $u'$ ).*
  - (iii) *Output the set  $\{u_1, \dots, u_{d_v}\} \setminus \{v\}$ .*

## 4 Pseudorandomness of the Main Construction

This section presents our main result (Theorem 3) that establishes the pseudorandomness of construction 1. It is proved by reducing it to our main technical result (Theorem 3.1) which asserts a negligible distance between views produced by the  $G_{\text{IBin}}$  and the  $G(N, p)$  models.

**Theorem 3 (Pseudorandomness of construction 1)** *Let  $\{p_n\}_{n \in \mathbb{N}}$  be arbitrary proper densities. Then, assuming the existence of one-way functions, the graphs distributions  $\{\mathcal{G}_n\}_{n \in \mathbb{N}}$  produced by construction 1 on inputs  $(1^n, p_n)$  are neighborhood-queries pseudorandom w.r.t.  $\{G(N, p_n)\}_{n \in \mathbb{N}}$ .*

**Proof (Theorem 3).** Correctness is trivial for  $p_n \leq N^{-1.5}$ , since the views of a  $G(N, p)$  graph reveal adjacencies only with negligible probability. For larger densities, by Theorems 1 and 2 the interval-sum function sub-routine and the involution sub-routine used by our construction are efficient. Since the remaining computations executed by our construction are trivial, the entire construction is efficient. Pseudorandomness - Using the transitivity of indistinguishability, we establish the pseudorandomness of  $\mathcal{G}_n$  by demonstrating (i) the indistinguishability of  $\mathcal{G}_n$  from  $G_{\text{IBin}}(N, p_n)$ , and (ii) the indistinguishability of  $G_{\text{IBin}}(N, p_n)$  from  $G(N, p_n)$ . Part (i) follows by a standard argument from the pseudorandomness of both the interval summable functions and the involutions used. Indeed, if (i) was false there would exist an efficient distinguisher  $D$  between  $\mathcal{G}_n$  and  $G_{\text{IBin}}(N, p_n)$ . This means that by combining our construction with  $D$ , one gets an efficient procedure for distinguishing between truly random interval summable functions and involutions and the pseudorandom ones - a contradiction to their pseudorandomness. Part (ii) follows immediately from Theorem 3.1, since the statistical distance between views produced by  $G_{\text{IBin}}$  and  $G(N, p)$  upper-bounds the distinguishing advantage of the distinguisher. ■

**Theorem 3.1 (Statistical distance between views produced by  $G_{\text{IBin}}$  and  $G(N, p)$ )** *Let  $C$  be an efficient distinguisher and let  $\underline{V}, \bar{V}$  denote the distributions of the view of  $C$  as the input graphs are sampled from either  $G_{\text{IBin}}(N, p_n)$  or  $G(N, p_n)$ , respectively. Then, the statistical distance between  $\underline{V}$  and  $\bar{V}$  is negligible.*

**Proof (Theorem 3.1).** Fix  $n$  and assume w.l.o.g. that the distinguisher  $C$  is a circuit.<sup>6</sup> We may further assume  $C$  to be deterministic, as the coin-tosses that produce the largest statistical distance between the views  $\underline{V}, \bar{V}$  can be hard-wired into the circuit. Let  $v_1, \dots, v_q$  denote the vertex-queries of  $C$ , and let  $R_1, \dots, R_q$  denote the responses that  $C$  receives. Thus  $R_j$  is the entire neighborhood of  $v_j$ , and  $\vec{R} = \{R_1, \dots, R_q\}$  denotes the entire view. Note that  $v_j, R_j, \vec{R}$  are all random variables as probabilities are taken over the choice of the graph (from either  $G_{\text{IBin}}$  or  $G(N, p)$ ). Next, let  $u_j, \Gamma_j$  and  $\vec{\Gamma}$ , respectively denote specific values of  $v_j, R_j$  and  $\vec{R}$ . As  $\overline{\text{Pr}}[\cdot], \underline{\text{Pr}}[\cdot]$  denote probabilities taken over  $G(N, p_n)$  and  $G_{\text{IBin}}(N, p_n)$ , respectively, our goal is to establish a negligible upper bound on

$$\Sigma_{\vec{\Gamma}} \left| \overline{\text{Pr}} \left[ \vec{R} = \vec{\Gamma} \right] - \underline{\text{Pr}} \left[ \vec{R} = \vec{\Gamma} \right] \right|. \quad (1)$$

**The proof proceeds as follows:** (i) We first separate the sum in equation 1 into ‘likely’ terms and into terms with either an unlikely port sequence or an unlikely view (formal definitions are given in the next paragraph). (ii) Next, the (un-surprising) negligible bound on the contribution of the unlikely terms is claimed in Lemma 2 and proved in Appendix 2. (iii) Then, the negligible bound for the likely terms is claimed in Lemma 1. Lemma 1 (which considers the entire view) is proved by reducing it to Claim 1 which considers the distributions of

<sup>6</sup>we thus strengthen the distinguisher *without* assuming our OWFs to be hard to invert even for circuits (and not only for Turing machines).

the next reply in the view (given the previous replies) and establishes sufficient closeness of these distributions in the  $G_{\text{IBin}}$  and the  $G(N, p)$  models. (iv) Finally, claim 1 itself is proved by first observing that in the  $G(N, p)$  case the size of the next neighborhood  $\Gamma_j$  has Binomial distribution, with each specific vertex being equally likely to appear in  $\Gamma_j$ . Thus, the main technical part of the proof is to analyze the distribution of the next reply in the  $G_{\text{IBin}}$  model. It will turn out that some terms concerning the distribution of the random port sequence will cancel nicely with other terms concerning the distribution of the random matching (given the port sequence). This way an (almost) Binomial distribution is established for the size of the next reply in the  $G_{\text{IBin}}$  case and claim 1 follows.

• **Definitions and notation.** An improper edge is either a self-edge, a multi-edge or an edge connected to the parity-vertex. A port is called proper if the edge it induces is proper, and a vertex is proper if all its ports are. A degree  $t_v$  is likely if  $0 \leq t_v \leq d_{\max}$  for  $d_{\max} = \lg N \lceil p(N-1) \rceil$ . An entire port sequence  $\vec{t}$  is likely if all the degrees  $t_v$  are likely and in addition  $\sum_{v=1}^N t_v = \mu(1 \pm \epsilon)$ , where  $\mu = N(N-1)p$  is the expected value of the sum, and the error-term is  $\epsilon = \frac{\lg \lg N}{\sqrt{pN}}$ . We let  $\mathbb{D}$  denote the set of all likely port sequences. The random variable that indicates the resulting port sequence is denoted  $\vec{d}$ .

A view  $\vec{\Gamma}$  is improper if some query-vertex  $v_j$  is improper. We let  $\mathbb{V}$  denote the collection of all ‘likely’ views that are simultaneously (i) Reply-collision-free - for any  $i < j$ ,  $[\Gamma(v_i) \cap \Gamma(v_j)] \setminus \{v_1, \dots, v_j\} = \emptyset$ . Namely, the distinguisher detects no ‘non-trivial’ collisions, but may produce trivial collisions by choosing, say, some  $v_2 \in \Gamma(v_1)$ , and some  $v_3 \in \Gamma(v_2)$  so  $v_2 \in [\Gamma(v_1) \cap \Gamma(v_3)]$ . (ii) Have small neighbor-sets -  $|\Gamma_j| \leq d_{\max} = \lg N \lceil p(N-1) \rceil$  for all  $j$ . (iii) Contain only proper vertices (as defined above). Given that a (partial) view  $\{\Gamma_1, \dots, \Gamma_{j-1}\}$  is likely, we say that a following reply  $\Gamma_j$  is likely if  $\{\Gamma_1, \dots, \Gamma_j\}$  remains likely.

**Separating the likely and unlikely terms.** The triangle inequality gives

$$\begin{aligned} & \left| \Sigma_{\vec{\Gamma}} \left[ \overline{\Pr} \left[ \vec{R} = \vec{\Gamma} \right] - \underline{\Pr} \left[ \vec{R} = \vec{\Gamma} \right] \right] \right| = \\ & \Sigma_{\vec{\Gamma} \in \mathbb{V}} \left| \overline{\Pr} \left[ \vec{R} = \vec{\Gamma} \right] - \left( \underline{\Pr} \left[ \vec{R} = \vec{\Gamma}, \vec{d} \in \mathbb{D} \right] + \underline{\Pr} \left[ \vec{R} = \vec{\Gamma}, \vec{d} \notin \mathbb{D} \right] \right) \right| + \\ & \Sigma_{\vec{\Gamma} \notin \mathbb{V}} \left| \overline{\Pr} \left[ \vec{R} = \vec{\Gamma} \right] - \underline{\Pr} \left[ \vec{R} = \vec{\Gamma} \right] \right| \leq \\ & \underbrace{\Sigma_{\vec{\Gamma} \in \mathbb{V}} \left| \overline{\Pr} \left[ \vec{R} = \vec{\Gamma} \right] - \underline{\Pr} \left[ \vec{R} = \vec{\Gamma}, \vec{d} \in \mathbb{D} \right] \right|}_{\stackrel{\text{def}}{=} T_1} + \\ & \underbrace{\underline{\Pr} \left[ \vec{R} \in \mathbb{V}, \vec{d} \notin \mathbb{D} \right] + \overline{\Pr} \left[ \vec{R} \notin \mathbb{V} \right] + \underline{\Pr} \left[ \vec{R} \notin \mathbb{V} \right]}_{\stackrel{\text{def}}{=} T_2} \end{aligned}$$

(this a-symmetric separation of events is crucial to our argument). Thus, proving Theorem 3.1 reduces to establishing the following lemmata.

**Lemma 1 (Statistical distance between *likely* views)** For  $C, \overline{\text{Pr}}[\cdot], \underline{\text{Pr}}[\cdot], \vec{R}, \mathbb{V}, \mathbb{D}$  and  $T_1$  as above the term  $T_1$  is negligible.

**Lemma 2 (Statistical distance between *unlikely* views)** For  $C, \overline{\text{Pr}}[\cdot], \underline{\text{Pr}}[\cdot], \vec{R}, \mathbb{V}, \mathbb{D}$ , and  $T_2$  as above the term  $T_2$  is negligible.

We prove Lemma 2 in Appendix 2 (section 6.2) and continue with Lemma 1.

**Proof of Lemma 1.** Assume w.l.o.g. that  $C$  performs the same number of queries  $q$  on any  $N$ -vertex graph. Then for any likely view  $\vec{\Gamma} = \{\Gamma_1, \dots, \Gamma_q\} \in \mathbb{V}$  we have

$$\begin{aligned} \overline{P} &\stackrel{\text{def}}{=} \overline{\text{Pr}} \left[ \vec{R} = \vec{\Gamma} \right] = \prod_{j=1}^q \overline{P}_j, \\ \underline{P} &\stackrel{\text{def}}{=} \underline{\text{Pr}} \left[ \vec{R} = \vec{\Gamma}, \vec{d} \in \mathbb{D} \right] = \underline{\text{Pr}} \left[ \vec{d} \in \mathbb{D} \right] \prod_{j=1}^q \underline{P}_j, \end{aligned}$$

where

$$\overline{P}_j = \overline{\text{Pr}} [R_j = \Gamma_j \mid R_1 = \Gamma_1, \dots, R_{j-1} = \Gamma_{j-1}]$$

and

$$\underline{P}_j = \underline{\text{Pr}} \left[ R_j = \Gamma_j \mid \vec{d} \in \mathbb{D}, R_1 = \Gamma_1, \dots, R_{j-1} = \Gamma_{j-1} \right].$$

We will show (Claim 1) that for some negligible  $\epsilon$ , and any  $\vec{\Gamma} \in \mathbb{V}$  and  $0 \leq j \leq q$  we have  $\underline{P}_j = \overline{P}_j(1 \pm \epsilon)$ , as well as  $\underline{\text{Pr}} \left[ \vec{d} \in \mathbb{D} \right] \geq 1 - \epsilon$ . As  $q$  is *poly*( $n$ )-bounded and  $\epsilon$  is negligible, then  $q\epsilon = o(1)$  so we get

$$\underline{P} = \overline{P}(1 \pm \epsilon)^{q+1} = \overline{P}(1 \pm \Theta(\epsilon q)).$$

Consequently,

$$T_1 = \sum_{\vec{\Gamma} \in \mathbb{V}} \Theta(\epsilon q) \overline{\text{Pr}} \left[ \vec{R} = \vec{\Gamma} \right] = \Theta(\epsilon q) \sum_{\vec{\Gamma} \in \mathbb{V}} \overline{\text{Pr}} \left[ \vec{R} = \vec{\Gamma} \right] = \Theta(\epsilon q),$$

which is negligible. Therefore Lemma 1 follows once we establish claim 1.  $\blacksquare$

**Claim 1** For  $\overline{P}_j, \underline{P}_j$ , and  $\mathbb{D}$  as above there exist a negligible  $\epsilon$ , s.t. for all likely views  $\vec{\Gamma} \in \mathbb{V}$  and all  $0 \leq j \leq q$  then  $\underline{P}_j = \overline{P}_j(1 \pm \epsilon)$ , and  $\underline{\text{Pr}} \left[ \vec{d} \in \mathbb{D} \right] \geq 1 - \epsilon$ .

**Proof (Claim 1).** We focus on the  $\underline{P}_j = \overline{P}_j(1 \pm \epsilon)$  part, as the  $\underline{\text{Pr}} \left[ \vec{d} \in \mathbb{D} \right] \geq 1 - \epsilon$  part merely states that the port sequence is likely (and is proved in Appendix 2, claim 3).

• **Notation.** Let  $W$  denote the set of vertices  $w \notin \left( \left( \bigcup_{i=1}^{j-1} \Gamma_i \right) \cup \{v_1, \dots, v_j\} \right)$  that haven't appeared in the view up to stage  $j$ , and let  $N' = |W|$ . Let  $A$  denote the event that the next reply  $\Gamma_j$  is likely (Thus  $\Gamma_j$  has size  $\leq d_{\max}$ , it intersects no previous  $\Gamma_i$ , and the query-vertex  $v_j$  is proper). Consider an arbitrary likely degree  $k \leq d_{\max}$ , and let  $H$  denote the event that a specific (partial) view has been observed, namely that  $(R_1 = \Gamma_1, \dots, R_{j-1} = \Gamma_{j-1})$ . Similarly, let  $H'$  denote the event  $(\vec{d} \in \mathbb{D}) \cap (R_1 = \Gamma_1, \dots, R_{j-1} = \Gamma_{j-1})$ .

As the claim deals only with likely views, then the view is reply-collision-free (see notations at page 8) so the next query vertex  $v_j$  appears in either one or none of the previous replies  $\Gamma_1, \dots, \Gamma_{j-1}$ . Assume w.l.o.g. that the first case holds (adapting the proof to the complement case is trivial). Therefore, whenever  $A$  occurs, we have:  $(|R_j| = k)$  iff  $v_j$  has precisely  $k - 1$  neighbors in  $W$ .

**The  $\mathbf{G(N, p_n)}$  case (Claim 1).** By the above and by the total independence of edges in  $G(N, p_n)$ ,

$$\overline{\Pr}[A, |R_j| = k | H] = \binom{N'}{k-1} p^{k-1} (1-p)^{N'-(k-1)}$$

with all specific choices of vertices  $w_1, \dots, w_{k-1} \in W$  for  $R_j$  being equiprobable.

**The  $\mathbf{G_{\Pi Bin}}$  case (Claim 1).** It isn't too hard to see (see Appendix 2, Claim 2 for formal proof) that whenever  $A$  holds, the symmetry of the  $G_{\Pi Bin}$  model implies that all specific choices of vertices  $w_1, \dots, w_{k-1} \in W$  are equiprobable for  $R_j$  (just as in  $G(N, p_n)$ ). Thus, it remains to show that (analogously to the  $G(N, p)$  case) the next reply has approximately Binomially distributed size, that is, to prove the following claim:

**Claim 1.1 (Establishing the (approximately) Binomial size of the next reply in the  $\mathbf{G_{\Pi Bin}}$  case.)** For  $A, H', N'$  as above,

$$\underline{\Pr}[A, |R_j| = k | H'] \sim \binom{N'}{k-1} p^{k-1} (1-p)^{N'-(k-1)}.$$

**Proof.** The following intuitive argument and the formal proof given later will apply similar ideas.

**The informal argument.** Let  $d_v^* = |\Gamma(v_j) \cap W|$ . Again, since  $v_j$  appears in the current view (that is, in  $\{\Gamma_1, \dots, \Gamma_{j-1}\}$ ) precisely once, then whenever  $A$  occurs, we have:  $(|R_j| = k)$  iff  $v_j$  has precisely  $k - 1$  neighbors in  $W$ . Hence,

$$\begin{aligned} \underline{\Pr}[A, |R_j| = k | H'] &= \underline{\Pr}[A, d_v^* = k - 1 | H'] \\ &= \frac{\underline{\Pr}[A, d_v^* = k - 1, H']}{\underline{\Pr}[H']} \\ &= \frac{\underline{\Pr}[A | H', d_v^* = k - 1] \underline{\Pr}[H' | d_v^* = k - 1] \underline{\Pr}[d_v^* = k - 1]}{\sum_t \underline{\Pr}[H' | d_v^* = t] \underline{\Pr}[d_v^* = t]}, \end{aligned}$$

(the sum taken over all likely degrees  $t$ ).

We will soon argue that  $\underline{\Pr}[H' | d_v^* = t'] \sim \underline{\Pr}[H' | d_v^* = t'']$  holds for any pair of likely degrees  $t', t''$ . Assuming this, we may cancel out (up to negligible terms) the  $\underline{\Pr}[H' | d_v^* = k - 1]$  from the nominator with the terms  $\underline{\Pr}[H' | d_v^* = t]$  from the denominator. Since the event  $A$  is extremely likely, then the term  $\underline{\Pr}[A | H', d_v^* = k - 1] \sim 1$  and cancels too. Thus,

$$\underline{\Pr}[A, |R_j| = k | H'] \sim \frac{\underline{\Pr}[d_v^* = k - 1]}{\sum_t \underline{\Pr}[d_v^* = t]} \sim \underline{\Pr}[d_v^* = k - 1],$$

and our claim follows (here  $\sum_t \Pr[d_v^* = t] \sim 1$  as we sum over all likely degrees).

To demonstrate that  $\Pr[H' | d_v^* = t'] \sim \Pr[H' | d_v^* = t'']$  (this is where our argument becomes informal), we first assume that all the degrees except  $d_{v_j}$  are fixed. We consider the equivalent model  $G_{\text{IBin}}^{\text{OTF}}$ , where the view  $H$  is produced by repeatedly matching each of the available ports  $\sigma_i$  of the current query vertex with a random available port. Let  $\sigma_1, \dots, \sigma_z$  denote the ports of the first  $j - 1$  query vertices. Clearly, when  $\sigma_i$  is matched, any available port  $\tau$  is chosen with probability precisely  $1/E_i$  for  $E_i = (\sum_v d_v) + 1 - 2i$ . Thus, the difference between the cases  $d_v = t'$  and  $d_v = t''$  is that each choice (of matching some  $\tau$  with  $\sigma_i$ ) occurs w.p.  $1/E_i'$  instead of  $1/E_i''$ . Since  $E_i', E_i''$  are both super-polynomial (as the port sequence is likely), and since  $E_i' - E_i'' = t' - t'' \leq \text{poly}(n)$ , this induces an insignificant difference between the resulting distributions. This difference remains insignificant even as diversities accumulate over the  $\text{poly}(n)$  many ports  $\sigma_i$  that are matched to decide  $H'$ . Finally, as this holds for any choice of the degrees (excluding  $d_{v_j}$ ), one can get  $\Pr[H' | d_v^* = t'] \sim \Pr[H' | d_v^* = t'']$ .

The formal proof of Claim 1.1 appears in Appendix 1. Given Claim 1.1, then Claim 1 follows, so Lemma 1 and hence our main Theorems 3.1 and 3 follow as well. ■

## Acknowledgements.

We thank Gil Segev, Noam Livne and the anonymous referees for carefully reading and commenting on a draft of this paper.

## References

- [1] N. Alon, J. Spencer. *The Probabilistic Method*. John Wiley, New York, 1992.
- [2] B. Bollobás. *A probabilistic proof of an asymptotic formula for the number of labelled regular graphs*. Preprint Series, Matematisk Institut, Aarhus Universitet, 1979.
- [3] J. Black, P. Rogaway. *Ciphers with Arbitrary Finite Domains*. CT-RSA 2002: 114-130.
- [4] P. Erdős and A. Rényi. *On random graphs I*. Publicationes Mathematicae 6 (1959), 290–297.
- [5] P. Erdős and A. Rényi. *On the evolution of random graphs*. Publications of the Mathematical Institute of the Hungarian Academy of Sciences, 5:17–61, 1960.
- [6] A. Gilbert, S. Guha, P. Indyk, Y. Kotidis, S. Muthukrishnan, and M. Strauss. *Fast, Small-Space Algorithms for Approximate Histogram Maintenance*. In 34<sup>th</sup> annual ACM symposium on Theory of computing, 389-398, 2002.
- [7] O. Goldreich, S. Goldwasser, S. Micali. *How to Construct Random Functions*. Journal of the ACM, vol. 33, no. 4, 276–288, 1985.
- [8] O. Goldreich, S. Goldwasser, A. Nussboim. *On the Implementation of Huge Random Objects*. In proc. 44<sup>th</sup> IEEE Symp. on Foundations of Computer Science, 68–79, 2003.

- [9] O. Goldreich, D. Ron. *Property Testing in Bounded Degree Graphs*. In Proceedings of the 29<sup>th</sup> ACM Symposium on Theory of Computing, pages 406-415, 1997.
- [10] E. Kaplan, M. Naor, O. Reingold. *Derandomized Constructions of  $k$ -Wise (Almost) Independent Permutations*. APPROX-RANDOM 2005, 354-365.
- [11] M. Luby and C. Rackoff. *How to Construct Pseudo-Random Permutations from Pseudo-Random Functions*. In SIAM J. on Computing, Vol. 17, 373-386, 1998.
- [12] M. Naor, A. Nussboim. *Implementing Huge Sparse Random Graphs*. Available at <http://www.wisdom.weizmann.ac.il/~asafn/PAPERS/sparseGnp.ps>.
- [13] M. Naor, A. Nussboim, E. Tromer. *Efficiently Constructible Huge Graphs that Preserve First Order Properties of Random Graphs*. Proceedings of the 2<sup>nd</sup> Theory of Cryptography Conference, 66-85, 2005.
- [14] M. Naor and O. Reingold. *Constructing Pseudo-Random Permutations with a Prescribed Cyclic Structure*. In Journal of Crypto. Vol. 15(2), pages 97-102, 2002.
- [15] M. Naor and O. Reingold. *On the construction of pseudorandom permutations: Luby-Rackoff revisited*. J. of Cryptology, 1999. Preliminary Version: STOC 1997.

## 5 Appendix 1: Proving the $G_{\text{IBin}}$ Case of Claim 1.

The entire Appendix 1 provides a formal proof of Claim 1.1.

Let  $\Phi = \Pr[|R_j| = k, A, H']$  and  $\Psi = \Pr[H']$ . We have  $\Pr[|R_j| = k, A | H'] = \Phi/\Psi$ . Recall that  $\vec{d} = (d_1, \dots, d_N)$  is the random variable that denotes the port sequence and let  $\vec{t} = (t_1, \dots, t_N)$  denote specific values of  $\vec{d}$ . One can expand both  $\Phi$  and  $\Psi$  by summing over all possible port sequences:  $\Phi = \sum_{\vec{t}} \Pr[\vec{d} = \vec{t}, |R_j| = k, A, H']$  and  $\Psi = \sum_{\vec{t}} \Pr[\vec{d} = \vec{t}, H']$ . Our argument will exploit a slightly different expansion where (i) we ignore the unlikely port sequences  $\vec{t}$  (their contribution is null since the event  $H'$  means that the port sequence is likely) (ii) we first sum over all possible likely degrees for the vertices that appear in previous replies, and only then sum over all possible likely degrees for the remaining vertices. The second sum will be taken only over degrees that complete the previous ones into a likely port sequence that is consistent with the current view. This separation of the sum is simple, yet, requires the following notation.

- **Further notation.** Let  $U = \{u_1, \dots, u_{|U|}\} = \left(\bigcup_{i=1}^{j-1} \Gamma_i\right) \setminus \{v_1, \dots, v_j\}$  denote the set of non-query vertices that appear in the first  $j-1$  replies. Let  $\vec{c} = (c_1, \dots, c_{|U|})$  indicate the number of ports each  $u \in U$  possesses. Let  $\vec{d}', \vec{t}' \in \{0, \dots, N-1\}^{N-|U|-1}$  denote the sub-sequences derived from either  $\vec{d}$  or  $\vec{t}$  (respectively) by omitting the  $|U|+1$  coordinates that correspond to  $v_j$  and to all  $u \in U$ . We also wish to consider the inverse operation of completing a sub-sequence  $\vec{t}'$  into a complete port sequence with the  $|U|+1$  new coordinates that indicate that  $d_{v_j} = \ell$  and that  $d_{u_i} = c_i$  (for  $i = 1, \dots, |U|$ ). The resulting complete port sequence is denoted  $\vec{t}(\ell, \vec{c})$ . Next, let  $\mathbb{C} = \{\vec{c} \mid 1 \leq c_1, \dots, c_{|U|} \leq d_{\max}\}$  denote the collection of vectors that assign a degree only to vertices  $u \in U$  s.t. all the degrees are likely. (we require  $1 \leq c_i$  instead of  $0 \leq c_i$  since  $u_i$  appears in some  $\Gamma_{i'}$ ).

Finally, given likely degrees  $0 \leq \ell \leq d_{\max}$ , and  $\vec{c} \in \mathbb{C}$  let  $\mathbb{D}_{\ell, \vec{c}}$  denote the set of all sub-sequences that (i) are consistent with all the cardinalities  $|\Gamma_i|$  in the current view, and (ii) produce a likely sequence when  $d_{v_j} = \ell, d_{u_1} = c_1, \dots, d_{u_{|U|}} = c_{|U|}$ . Namely,  $\mathbb{D}_{\ell, \vec{c}} = \{\vec{t}' \mid (\bigwedge_{i=1}^{j-1} t_{v_i} = |\Gamma_i|), \vec{t}(\ell, \vec{c}) \in \mathbb{D}\}$ . We can now expand,

$$\begin{aligned} \Phi &= \sum_{\vec{c} \in \mathbb{C}} \sum_{\vec{t}' \in \mathbb{D}_{k, \vec{c}}} \underline{\Pr} \left[ |R_j| = k, A, H', \vec{d} = \vec{t}(k, \vec{c}) \right] \\ &= \sum_{\vec{c} \in \mathbb{C}} \sum_{\vec{t}' \in \mathbb{D}_{k, \vec{c}}} \underline{\Pr} \left[ \vec{d} = \vec{t}(k, \vec{c}) \right] \underline{\Pr} \left[ H', A \mid \vec{d} = \vec{t}(k, \vec{c}) \right] \end{aligned}$$

and,

$$\begin{aligned} \Psi &= \sum_{\ell=1}^{d_{\max}} \sum_{\vec{c} \in \mathbb{C}} \sum_{\vec{t}' \in \mathbb{D}_{\ell, \vec{c}}} \underline{\Pr} \left[ H', \vec{d} = \vec{t}(\ell, \vec{c}) \right] \\ &= \sum_{\ell=1}^{d_{\max}} \sum_{\vec{c} \in \mathbb{C}} \sum_{\vec{t}' \in \mathbb{D}_{\ell, \vec{c}}} \underline{\Pr} \left[ \vec{d} = \vec{t}(\ell, \vec{c}) \right] \underline{\Pr} \left[ H' \mid \vec{d} = \vec{t}(\ell, \vec{c}) \right]. \end{aligned}$$

The latter expressions are now analyzed using Facts 1 – 3 (which are proved later) to derive the desired  $\Phi/\Psi \sim \frac{N^{k-1}}{(k-1)!} p^{k-1} (1-p)^N$  approximation. Facts 1 – 3 enable us to cancel out (up to  $\sim$ ) some terms in the nominator  $\Phi$  with some terms in the denominator  $\Psi$ . The substantial counting arguments are captured by Fact 1 which shows that given the degrees of the first  $j$  query vertices and the degrees of the vertices in the first  $j-1$  replies - then the distribution of the next reply is roughly independent of the degrees of the vertices that haven't appeared in the view so far (as long as the port sequence is likely). The unsurprising Fact 2 merely establishes the closure of  $\sim$  under composition of arithmetic operations and is proved in Appendix 2 (section 6.4). The unsurprising Fact 3 demonstrates that even as we condition on several specific degrees (of the current query vertex  $v_j$  and of all previous reply vertices) we still get a likely port sequence with overwhelming probability. It is proved at the end of Appendix 2 section 6.2.

**Fact 1 (distribution of views given parts of the port sequence)** *Let  $0 \leq t_{i_1}, \dots, t_{i_{j-1}} \leq d_{\max}$  be arbitrary likely degrees and let  $D \subseteq \mathbb{D}$  denote the collection of all likely port sequences with  $d_{v_1} = t_{i_1}, \dots, d_{v_{j-1}} = t_{i_{j-1}}$ . Then there exists a value  $\Lambda = \Lambda(t_{i_1}, \dots, t_{i_{j-1}})$  and a negligible  $\delta$  s.t. for any port sequence  $\vec{t} \in D$ , both  $\underline{\Pr} \left[ H', A \mid \vec{d} = \vec{t} \right], \underline{\Pr} \left[ H' \mid \vec{d} = \vec{t} \right] = \Lambda[\prod_{u \in U} d_u] d_{v_j} (1 \pm \delta)$ .*

**Fact 2 (closure of  $\sim$ )** *Let  $A_i^r, \bar{A}_i^r, B_j^r, \bar{B}_j^r \geq 0$  for  $r = 1, 2, i = 1, \dots, m_i, j = 1, \dots, m_j$  (the  $r$  is merely an index not a power). If there exists a negligible  $\epsilon$  s.t.  $A_i^r = \bar{A}_i^r (1 \pm \epsilon), B_j^r = \bar{B}_j^r (1 \pm \epsilon)$ , for all  $i, j, r$  then  $\left( \sum_{i=1}^{m_i} A_i^1 A_i^2 / \sum_{j=1}^{m_j} B_j^1 B_j^2 \right) \sim \left( \sum_{i=1}^{m_i} \bar{A}_i^1 \bar{A}_i^2 / \sum_{j=1}^{m_j} \bar{B}_j^1 \bar{B}_j^2 \right)$ .*

**Fact 3 (probability of likely port sequence given a partial port sequence)** For any likely degree  $\ell$  and any  $\vec{c} \in \mathbb{C}$ ,  $\sum_{\vec{t}' \in \mathbb{D}_{\ell, \vec{c}}} \underline{\Pr} [\vec{d}' = \vec{t}'] \sim 1$ .

**Concluding Claim 1.1 (via Facts 1 – 3).** Applying Fact 1 and then Fact 2 gives  $\Phi/\Psi \sim \Phi'/\Psi'$  for some  $\Lambda$  and for

$$\begin{aligned}\Phi' &= \sum_{\vec{c} \in \mathbb{C}} \sum_{\vec{t}' \in \mathbb{D}_{k, \vec{c}}} \underline{\Pr} [\vec{d} = \vec{t}'(k, \vec{c})] \Lambda [\Pi_{u \in U} d_u] k \\ \Psi' &= \sum_{\ell=1}^{d_{\max}} \sum_{\vec{c} \in \mathbb{C}} \sum_{\vec{t}' \in \mathbb{D}_{\ell, \vec{c}}} \underline{\Pr} [\vec{d} = \vec{t}'(\ell, \vec{c})] \Lambda [\Pi_{u \in U} d_u] \ell.\end{aligned}$$

As  $\Lambda$  cancels out, the total independence of degrees gives

$$\begin{aligned}\Phi' &= \sum_{\vec{c} \in \mathbb{C}} \sum_{\vec{t}' \in \mathbb{D}_{k, \vec{c}}} \underline{\Pr} [d_{v_j} = k] (\Pi_{u \in U} \underline{\Pr} [d_u = c_u]) \underline{\Pr} [\vec{d}' = \vec{t}'] (\Pi_{u \in U} d_u) k, \\ \Psi' &= \sum_{\ell=1}^{d_{\max}} \sum_{\vec{c} \in \mathbb{C}} \sum_{\vec{t}' \in \mathbb{D}_{\ell, \vec{c}}} \underline{\Pr} [d_{v_j} = \ell] (\Pi_{u \in U} \underline{\Pr} [d_u = c_u]) \underline{\Pr} [\vec{d}' = \vec{t}'] (\Pi_{u \in U} d_u) \ell.\end{aligned}$$

Changing the order of summation gives,

$$\begin{aligned}\Phi' &= \underline{\Pr} [d_{v_j} = k] k \sum_{\vec{c} \in \mathbb{C}} \Pi_{u \in U} (d_u \underline{\Pr} [d_u = c_u]) \sum_{\vec{t}' \in \mathbb{D}_{k, \vec{c}}} \underline{\Pr} [\vec{d}' = \vec{t}'], \\ \Psi' &= \sum_{\ell=1}^{d_{\max}} \underline{\Pr} [d_{v_j} = \ell] \ell \sum_{\vec{c} \in \mathbb{C}} \Pi_{u \in U} (d_u \underline{\Pr} [d_u = c_u]) \sum_{\vec{t}' \in \mathbb{D}_{\ell, \vec{c}}} \underline{\Pr} [\vec{d}' = \vec{t}'].\end{aligned}$$

Finally, let  $Q_\ell$  denote the probability of obtaining precisely  $\ell$  successes in  $N-1$  independent Bernoulli trials each with probability of success  $p$ . By Facts 3 and 2 the  $\sum_{\vec{t}' \in \mathbb{D}_{k, \vec{c}}} \underline{\Pr} [\vec{d}' = \vec{t}']$  and  $\sum_{\vec{t}' \in \mathbb{D}_{\ell, \vec{c}}} \underline{\Pr} [\vec{d}' = \vec{t}']$  in  $\Phi'/\Psi'$  cancel out (up to  $\sim$ ). Next,  $\sum_{\vec{c} \in \mathbb{C}} [\Pi_{u \in U} d_u \underline{\Pr} [d_u = c_u]]$  cancels out too. This justifies the first  $\sim$ :

$$\begin{aligned}\Phi'/\Psi' &\sim [k \underline{\Pr} [d_{v_j} = k]] / \left[ \sum_{\ell=1}^{d_{\max}} \ell \underline{\Pr} [d_{v_j} = \ell] \right] \\ &= \left[ k \binom{N-1}{k} p^k (1-p)^{N-1-k} \right] / \left[ \sum_{\ell=1}^{d_{\max}} \ell \binom{N-1}{\ell} p^\ell (1-p)^{N-1-\ell} \right].\end{aligned}$$

Next, as  $(1-p)^{N-1-k} \sim (1-p)^{N-1-(k-1)}$  and since for any  $poly(n)$ -bounded  $\ell$ ,  $\ell \binom{N-1}{\ell} = (N-\ell) \binom{N-1}{\ell-1} \sim N \binom{N-1}{\ell-1}$  (see fact 5 in Appendix 2), the first  $\sim$  holds

$$\Phi'/\Psi' \sim pN \left[ \binom{N-1}{k-1} p^{k-1} (1-p)^{N-1-(k-1)} \right] / pN \left[ \sum_{\ell=1}^{d_{\max}} \binom{N-1}{\ell-1} p^{\ell-1} (1-p)^{N-1-(\ell-1)} \right]$$

$$\begin{aligned}
&= \left[ Q_{k-1} / \sum_{\ell'=0}^{d_{\max}-1} Q_{\ell'} \right] \sim Q_{k-1} \\
&= \binom{N-1}{k-1} p^{k-1} (1-p)^{N-(k-1)} \\
&\sim \binom{N'}{k-1} p^{k-1} (1-p)^{N'-k+1}.
\end{aligned}$$

Here, the  $[\cdot] \sim Q_{k-1}$  is implied by  $\sum_{\ell'=0}^{d_{\max}-1} Q_{\ell'} \sim 1$  ( $\sum_{\ell' \geq d_{\max}} Q_{\ell'}$  is negligible as it implies an unlikely degree (Claim 3)). The final  $\sim$  is proved in Appendix 2 (Fact 4).

This shows that (analogously to the  $G(N, p)$  case) in the  $G_{\text{IBin}}$  case too the size of the next reply has almost Binomial distribution ( $\Pr[|R_j| = k, A | H'] \sim \binom{N'}{k-1} p^{k-1} (1-p)^{N'-k+1}$ ). Thus, completing the entire proof of Claim 1.1 (which implies Claim 1 and hence Lemma 1) reduces to establishing our main combinatorial Fact 1 (The much simpler Facts 2 and 3 are given in Appendix 2).

**Proof of Fact 1.** We compute the probability of a specific partial view  $H = (\Gamma_1, \dots, \Gamma_{j-1})$  given a specific likely port sequence  $\vec{t} = (t_1, \dots, t_N)$  (since the port sequence is likely the events  $H, H'$  are identical). As matchings are uniformly chosen at random, this probability equals the number of matchings  $M(H, \vec{t})$  consistent with both  $H$  and  $\vec{t}$ , divided by the number of matchings  $M(\vec{t})$  consistent only with  $\vec{t}$ . Now, as the view is likely (see section 4), then all non-query vertices  $u$  that arise in the view up to stage  $j-1$  appear only in a single reply  $\Gamma_i$ . Recall that  $U$  denotes the collection of such vertices  $u$ . Let  $S = \sum_{v=1}^N t_v$  denote the total number of ports, and let  $\bar{S} = S - \left( \sum_{i=1}^{j-1} t_{v_i} \right) - |U|$  denote the number of unmatched ports after  $j-1$  queries.

Throughout the proof we will use the equivalent model  $G_{\text{IBin}}^{\text{OTF}}$  instead of  $G_{\text{IBin}}$  (see section 2.2). In the  $G_{\text{IBin}}^{\text{OTF}}$  model, it clearly holds that  $M(H, \vec{t}) = M_{j-1}(H, \vec{t}) \cdot \overline{M}(\bar{S})$ , where  $M_{j-1}(H, \vec{t})$  is the number of ways to match only the ports of the query-vertices  $v_1, \dots, v_{j-1}$ , s.t. this sub-matching is consistent with both  $H$  and  $\vec{t}$ , and  $\overline{M}(\bar{S})$  denotes the number of possible matchings over  $\bar{S}$  elements. Therefore  $\overline{M}(\bar{S})$  counts the number of ways for completing a single matching of the ports of the query-vertices into a complete matching of all the ports.

Next, observe that  $M_{j-1}(H, \vec{t}) = \prod_{u \in U} t_u \prod_{i=1}^{j-1} [t_{v_i}!]$ . Indeed, fix  $\vec{t}$  and consider a specific sub-matching  $B$  over the ports of  $v_1, \dots, v_{j-1}$  that is consistent with  $H$ . Recall that arbitrary vertices  $w_1, w_2$  are connected by an edge iff some port of  $w_1$  is connected with some port of  $w_2$ . Thus, for any  $v_i$ , it holds that any of the  $t_{v_i}!$  possible ways of permuting the names of the ports of  $v_i$  yields a new sub-matching  $B'$  that induces the same edges and non-edges as  $B$  does. Similarly, as the view is likely, then for any  $u \in U$ , there exist a unique query-vertex  $v_i$ , that is matched with  $u$  via *some* port  $\rho_u$  of  $u$ . Again, any of the  $t_u$  choices of  $\rho_u$  induce the same adjacency pattern. All this holds independently for all  $v_i$  and all  $u \in U$ , so there are  $\prod_{u \in U} t_u \prod_{i=1}^{j-1} [t_{v_i}!]$  possible sub-matchings. Since the view is likely, there are no multi-edges, and hence these various sub-matchings

are indeed distinct. It is also easy to verify that all sub-matchings consistent with  $H$  can be produced as above, so  $M_{j-1}(H, \vec{t})$  is precisely  $\prod_{u \in U} t_u \prod_{i=1}^{j-1} [t_{v_i}]$ . To summarize, all the above implies,

$$\underline{\Pr} [H | \vec{t}] = M(H, \vec{t}) / M(\vec{t}) = \left( \prod_{i=1}^{j-1} t_{v_i}! \right) (\prod_{u \in U} t_u) \overline{M}(\overline{S}) / \overline{M}(S).$$

Therefore to complete the proof of Fact 1 it suffices provide some  $\Lambda$  and a negligible  $\delta$  s.t.  $\overline{M}(\overline{S}) / \overline{M}(S) = \Lambda(1 \pm \delta)$  regardless of the specific likely choice of  $\vec{t}$ . With  $m \stackrel{\text{def}}{=} S - \overline{S}$ , applying the well known formula  $\overline{M}(2x) = \frac{(2x)!}{x! 2^x}$  gives

$$\begin{aligned} \overline{M}(\overline{S}) / \overline{M}(S) &= (\overline{S})! (0.5S)! 2^{0.5S} \left[ S! (0.5\overline{S})! 2^{0.5\overline{S}} \right]^{-1} \\ &= \prod_{i=1}^m (S + 1 - 2i)^{-1} \\ &\quad \sim \prod_{i=1}^m S^{-1} \\ &= S^{-m} \sim (pN^2)^{-m}. \end{aligned}$$

Here, the second = follows as most terms in the nominator cancel with most terms in the denominator. The two concluding  $\sim$  are implied as the view and the port sequence are likely, so  $m \leq \text{poly}(n) \ll S \sim pN^2$  and the latter term is  $n^{\omega(1)}$  by the lower-bound on proper  $ps$ ; Specifically, standard calculations are used in Claim 6 (see Appendix 2) to obtain,

$$\prod_{i=1}^m (S + 1 - 2i)^{-1} = pN(N-1)^{-m} (1 \pm \delta)$$

for some negligible  $\delta$  which is independent of the specific value of  $S$  and  $m$ . Finally, as  $m, \delta$  are determined by the view  $H$  (regardless of the precise port sequence  $\vec{t}$ ), the  $\underline{\Pr} [H | \vec{t}]$  part of Fact 1 follows.

For the  $\underline{\Pr} [A, H | \vec{t}]$  part, recall  $A$  is the event that the next reply  $R_j$  keeps the view likely. It suffices to show that  $\Pr[A | H, \vec{t}] \geq 1 - \epsilon'$  for a single negligible choice of  $\epsilon'$  which is independent of the specific choice of a likely  $\vec{t}$ . This unsurprising result is established during the proof that the entire view is likely in Claim 4 part (ii).

This completes the entire proof of claim 1 so lemma 1 and hence our main Theorems 3.1 and 3 follow.  $\blacksquare$

## 6 Appendix 2: Proving Various Claims

### 6.1 Equiprobability of Various Vertices Appearing in a $G_{\text{PBIn}}$ View

We observe that in the  $G_{\text{PBIn}}$  model the reply-vertices are uniformly distributed among all non-query vertices (in contrast with the query-vertices  $v_j$  which are determined by the distinguisher). This observation immediately follows from the following claim.

**Claim 2** Consider a likely view in the  $G_{\text{IBin}}$  model, and recall that  $R_i$  is the random variable which denotes the  $i$ 'th reply, and that  $W$  denotes the set of vertices  $w \notin \left( \left( \bigcup_{i=1}^{j-1} \Gamma_i \right) \cup \{v_1, \dots, v_j\} \right)$  that haven't appeared in the view up to stage  $j$ . Then for any sequence of  $j-1$  likely replies  $\Gamma_1, \dots, \Gamma_{j-1}$  and any pair of possible next replies  $\Gamma_j = \{w_1 < \dots < w_k\}, \Gamma'_j = \{w'_1 < \dots < w'_k\} \subseteq W$  we have  $\Pr[R_j = \Gamma_j \mid R_1 = \Gamma_1, \dots, R_{j-1} = \Gamma_{j-1}] = \Pr[R_j = \Gamma'_j \mid R_1 = \Gamma_1, \dots, R_{j-1} = \Gamma_{j-1}]$ .

**Proof.** Consider the sub-configuration  $\bar{C}$  at stage  $j-1$ , defined by assigning a ports-number  $d_v$ , and ports  $\rho_{v,1}, \dots, \rho_{v,d_v}$  to each vertex  $v$ , and by some sub-matching defined over the ports of previous query vertices  $v_1, \dots, v_{j-1}$  (this sub-matching is a collection of  $\leq \sum_{i=1}^{j-1} d_{v_i}$  edges s.t. all the ports of the query-vertices are covered by the edges, and each edge touches at least a single port of a query-vertex). Let  $C(\Gamma_j), C(\Gamma'_j)$  denote the collection of all complete configurations (i.e. configurations that match all the ports) that are consistent with either  $\Gamma_j = \Gamma_j$  or with  $\Gamma_j = \Gamma'_j$ , respectively. It clearly suffices to provide a measure-preserving bijection  $\Phi : C(\Gamma_j) \rightarrow C(\Gamma'_j)$ .

The desired bijection replaces the  $w_r$ 's with the corresponding  $w'_r$ 's, as follows. We will define some permutation  $\pi$  over the entire vertex-set s.t. for any configuration  $C \in C(\Gamma_j)$ , then  $\Phi(C) = \Phi(C, \pi)$  where  $\Phi(C, \pi)$  is obtained from  $C$  by leaving the edges (that connect the ports) intact while renaming each port  $\rho_{v,i}$  as a port labeled by  $\rho_{\pi(v),i}$ . Given  $\Gamma_j, \Gamma'_j$  we define  $\pi = \pi(\Gamma_j, \Gamma'_j)$  by setting  $\pi(w_r) = w'_r$  for  $r = 1, \dots, k$ , and  $\pi(u) = u$  for  $u \notin \Gamma_j \cup \Gamma'_j$ . To define  $\pi$  over  $\Gamma'_j \setminus \Gamma_j$ , let  $\Gamma_j \setminus \Gamma'_j = \{w_{j_1} < \dots < w_{j_t}\}$  and  $\Gamma'_j \setminus \Gamma_j = \{w'_{j_1} < \dots < w'_{j_t}\}$  and set  $\pi(w'_{j_i}) = w_{j_i}$ .

To show that  $\Phi$  is a bijection, consider the dual permutation  $\pi' = \pi(\Gamma'_j, \Gamma_j)$  (the difference between  $\pi$  and  $\pi'$  is that the roles of  $\Gamma_j$  and  $\Gamma'_j$  is reversed). Note that  $\pi$  and  $\pi'$  are inverses. Indeed  $\pi(w_r) = w'_r$  and  $\pi'(w'_r) = w_r$ , while  $\pi(w'_{j_i}) = w_{j_i}$  and  $\pi'(w_{j_i}) = w'_{j_i}$ . Therefore, we let  $\Phi' : C(\Gamma'_j) \rightarrow C(\Gamma_j)$  be defined by  $\Phi'(C) = \Phi(C, \pi')$ . Since  $\pi, \pi'$  are inverses of each other then so are  $\Phi$  and  $\Phi'$ . Consequently,  $\Phi$  is a bijection.

Finally, for any configuration  $C \in C(\Gamma_j)$  the port sequence  $s$  of  $C$  and of  $\Phi(C)$  are equiprobable as they are the same up to permutation of coordinations. In addition, given the port sequence, all matchings are equiprobable. Thus,  $C$  and  $\Phi(C)$  are equiprobable. ■

## 6.2 Bounding the Probability of Unlikely Port-Sequences and Views

Claims 3 and 4 bound the probability of unlikely port sequences and views. Combined together they immediately imply Lemma 1.

**Claim 3** Recall that  $\mathbb{D}$  denotes the set of likely port sequences,  $\vec{d}$  is the random variable which indicates the actual port sequence, and  $\overline{\Pr}[\cdot], \underline{\Pr}[\cdot]$  denote probabilities taken over  $G(N, p)$  and  $G_{\text{IBin}}$ , respectively. Then,  $\overline{\Pr}[\vec{d} \notin \mathbb{D}], \underline{\Pr}[\vec{d} \notin \mathbb{D}]$  are both negligible.

**Proof.** We need to show that with overwhelming probability (i) the maximal degree (ports-number) among all vertices is bounded by  $\lg N \lceil p(N-1) \rceil$ , and (ii) the total number of ports falls in the range  $pN(N-1)(1 \pm \epsilon)$  for  $\epsilon = \frac{\lg \lg N}{\sqrt{p}N}$ .

For part (i) note that in both models ( $G(N, p)$  and  $G_{\text{IBin}}$ ), and for each vertex  $v$ , the degree  $d_v$  is the sum of  $M = N - 1$  independent Bernoulli trials, each with probability of success  $p$ . We note that  $\Pr[d_v = x]$  is decreasing for  $x \geq pM$  since

$$\begin{aligned} \frac{\Pr[d_v = x + 1]}{\Pr[d_v = x]} &= \frac{(M - x)p}{(x + 1)(1 - p)} \\ &< \frac{(M - Mp)p}{(Mp + 1)(1 - p)} \\ &< \frac{M(1 - p)p}{(Mp)(1 - p)} = 1, \end{aligned}$$

(the first  $<$  follows as the second expression is decreasing in  $x$ ). Thus for  $\Delta \stackrel{\text{def}}{=} \lg N \lceil pM \rceil$ , we have  $\Pr[d_v \geq \Delta] < M \Pr[d_v = \Delta] < M \binom{M}{\Delta} p^\Delta$ . Consequently,

$$\begin{aligned} \Pr[\exists v(d_v > \Delta)] &< NM \binom{M}{\Delta} p^\Delta \\ &< N^2 \left( \frac{eM}{\Delta} \right)^\Delta p^\Delta \\ &= N^2 \left( \frac{eMp}{\Delta} \right)^\Delta \\ &= N^2 \left( \frac{eMp}{\lg N \lceil pM \rceil} \right)^{\lg N \lceil pM \rceil} \\ &< N^2 \left( \frac{e}{\lg N} \right)^{\lg N} \\ &= N^2 \Theta(\lg N)^{-\lg N} \\ &= N^{\Theta(1)} N^{-\Omega(\lg \lg N)} \end{aligned}$$

which is negligible.

For part (ii), note that in the  $G_{\text{IBin}}$  model the total number of ports  $X \stackrel{\text{def}}{=} \sum_{v=1}^N d_v$  is the sum of  $M = N(N - 1)$  independent Bernoulli trials, each with probability of success  $p$ . By Chernoff's multiplicative bound  $\Pr[X \neq \mu(1 \pm \epsilon)] \leq e^{-\beta}$ , for  $\beta = \Theta(\epsilon^2 pM)$ . Since  $\epsilon = \frac{\lg \lg N}{\sqrt{p}N}$  then  $\epsilon^2 \geq \omega\left(\frac{\lg \lg N}{pN^2}\right)$  so  $e^{-\beta} = e^{-\omega(\lg \lg N)} = (\lg N)^{-\omega(1)}$  which is negligible too. A similar argument holds for the  $G(N, p)$  case. ■

**Claim 4** Recall that  $\mathbb{D}, \mathbb{V}$ , respectively denote the set of likely port sequences, and the set of likely views and that  $\vec{d}, \vec{R} = \{R_1, \dots, R_q\}$  are the random variables that respectively indicate the port sequence, and the view. Also Recall that  $\overline{\Pr}[\cdot], \underline{\Pr}[\cdot]$  denote probabilities taken over  $G(N, p)$  and  $G_{\text{IBin}}$ , respectively. Finally let  $B_j$  be the event that the  $j$ 'th reply is unlikely. Then,

1. In both models  $\overline{\Pr}[\vec{\Gamma} \notin \mathbb{V} \mid \vec{d} \in \mathbb{D}], \underline{\Pr}[\vec{\Gamma} \notin \mathbb{V} \mid \vec{d} \in \mathbb{D}]$  are both negligible.

2. In the  $G_{\text{IBin}}$  model, given any efficient distinguisher  $C$  there exists a negligible  $\epsilon_n$  s.t. for arbitrary likely replies  $\Gamma_1, \dots, \Gamma_{j-1}$  the following holds:  
 $\underline{\Pr}[B_j \mid R_1 = \Gamma_1, \dots, R_{j-1} = \Gamma_{j-1}] \leq \epsilon_n$ .

**Proof.** Let  $L_j$  denote the event that the current view  $\Gamma_1, \dots, \Gamma_j$  and the entire port sequence are both likely. We show that  $\Pr[\neg L_j \mid L_{j-1}]$  is negligible, and conclude that  $\Pr[\vec{\Gamma} \notin \mathbb{V} \mid \vec{d} \in \mathbb{D}] = \sum_{j=1}^q \Pr[\neg L_j \mid L_{j-1}]$  is negligible too (recall that the number of queries  $q$  is  $\text{poly}(\lg N)$  bounded). Notation: Recall that  $U = \left(\bigcup_{i=1}^{j-1} \Gamma_i\right) \setminus \{v_1, \dots, v_j\}$  denotes the set of non-query vertices that appear in the first  $j-1$  replies. Also recall that  $d_{\max} \stackrel{\text{def}}{=} \lg N[pN]$  is  $\leq \text{poly}(\lg N)$  by our original  $\frac{(\lg N)^{O(1)}}{N}$  upper bound on  $p$ .

**The  $G(N, p_n)$  case.** As all the degrees are likely  $|U| \leq qd_{\max} \leq \text{poly}(\lg N)$ . Note that  $\neg L_j$  holds iff  $v_j$  is connected to some  $u \in U$ . Hence the total independence of edges gives (via union bound)  $\overline{\Pr}[\neg L_j \mid L_{j-1}] = |U|p$  which is negligible as we handle only proper densities  $p = \frac{(\lg N)^{O(1)}}{N}$ .

**The  $G_{\text{IBin}}$  case.** We use the equivalent model  $G_{\text{IBin}}^{\text{OTF}}$  (see section 2.2) instead of  $G_{\text{IBin}}$ . Assume  $L_{j-1}$  holds and let  $\rho_{j_1}, \dots, \rho_{j_k}$  denote the ports of  $v_j$ . It suffices to consider the current step where  $\rho_{j_i}$  is matched with a random available port, and give a negligible bound on  $\underline{\Pr}[\rho_{j_i} \text{ is bad} \mid L_{j-1}]$ , so the claim will follow via a union-bound over the  $\text{poly}(\lg N)$  possible ports  $\rho_{j_i}$ . Here  $\rho_{j_i}$  is bad if it causes the view to be un-likely, namely if it is either improper or connects  $v_j$  to a vertex  $u \in U$ .

Let  $E = \sum_{v=i}^N d_v$  denote the total number of ports, and  $E'$  denote the number of currently available ports (at the moment when  $\rho_{j_i}$  is matched). Let  $I$  denote the number of available ports  $\rho$  that induce a bad edge (in case matched with  $\rho_{j_i}$ ), so  $\underline{\Pr}[\rho_{j_i} \text{ is bad}] = I/E'$ .

As the port sequence is likely,  $E \sim pN^2 = n^{\omega(1)}$  (by our original  $\frac{(n)^{\omega(1)}}{N^2}$  lower-bound on  $p$ ). Thus,  $E' \geq E - qd_{\max} = E - \text{poly}(n) \sim E = n^{\omega(1)}$ . Now  $I$  is  $\text{poly}(n)$  bounded. Indeed, there are at most  $d_{\max}(q-1)$  vertices in  $U$ , so these vertices are associated with at most  $d_{\max}^2(q-1)$  ports. In addition, at most  $d_{\max} - 1$  ports may create a self-loop or a multi-edge, and there is finally (possibly) a single parity-port. Thus  $I \leq d_{\max}^2 q \leq \text{poly}(n)$ , so  $\underline{\Pr}[\rho_{j_i} \text{ is bad}] = I/E'$  is negligible and part (i) follows. It is easy to verify that the latter negligible bound is independent of the specific choice of replies  $\Gamma_1, \dots, \Gamma_{j-1}$  so part (ii) follows too. ■

**Proof of Fact 3.** This fact ensures that with overwhelming probability we get a likely port sequence even after conditioning on arbitrary (likely) degrees for all previous reply vertices and for the current query vertex  $v_j$ .

**The proof.** If the port sequence is unlikely then either (i) some degree is unlikely or (ii) the total number of ports  $\sum_{v=1}^N d_v$  falls outside of the range  $S \stackrel{\text{def}}{=} pN(N-1)(1 \pm \epsilon)$  for  $\epsilon = \frac{\lg \lg N}{\sqrt{pN}}$ . Event (i) holds with negligible probability by Claim 3. When (ii) holds consider the sum  $S' \stackrel{\text{def}}{=} \sum_{v \notin U, v \neq v_j} d_v$  over  $N' = N - |U| - 1$  vertices. As  $|U|$  is  $\text{poly}(\lg N)$  bounded  $N' \sim N$ . Therefore, event (ii) gives  $S' \neq p(N')(N-1)(1 \pm \epsilon')$ , with  $\epsilon' = \Theta(\epsilon)$ . This event has negligible probability by the same argument as for showing that almost surely  $S = pN(N-1)(1 \pm \epsilon)$  (in Claim 3). ■

### 6.3 Pseudorandom Involutions over General Domains

We adapt the original [14] construction of pseudorandom involutions for domains of size  $2^n$  into domains of arbitrary 'proper' sizes, namely, any sequence of even sizes  $\{M_n\}_{n \in \mathbb{N}}$  where  $n^{\omega(1)} \leq M_n \leq 2^{\text{poly}(n)}$ . We start with the easy task of handling proper sizes  $M = M_n$  which are powers of 2, and based on that handle non powers of 2 later.

**Handling proper powers of 2.** We first recall that our pseudorandom graphs construction applies pseudorandom involutions (only) to match the set  $S$  of ports in the configurational model where  $|S| \leq N(N-1) = 2^{\Theta(n)}$  always holds and where only with negligible probability  $|S| \neq \mu(1 \pm \epsilon)$  for  $\mu = pN(N-1)$  and some negligible  $\epsilon$  (fact 3). Note that as we consider only proper densities  $p$  then  $\mu \geq n^{\omega(1)}$  so the total number of ports to match is super-polynomial.

Next, we recall that the [14] construction of pseudorandom involutions directly applies the standard Luby-Rackoff construction of pseudorandom permutations [11]. Thus the [14] construction inherits from [11] the properties of (i) achieving indistinguishability (up to negligible factors) from the truly random case whenever  $M \geq n^{\omega(1)}$  (ii) achieving efficiency whenever  $M \leq 2^{\text{poly}(n)}$ .<sup>7</sup>

**Handling proper non powers of 2.** As claimed before, we can provide a  $\text{poly}(n)$ -time (yet, not a very efficient) construction of pseudorandom involutions over arbitrary proper domains. Instead, we describe here a much more efficient and simpler construction. The simpler construction may fail to evaluate the value of the involution on some query element  $x$  (in such a case, the port associated with  $x$  in the configurational model is ignored). Failures, how-

<sup>7</sup>For readers familiar with the Luby Rackoff construction [11] as analyzed in [15] the latter claim is justified below. Part (ii) is trivial. For part (i) recall that the pseudorandomness proof relies only on (a) the pseudorandomness of the functions invoked in the second and third Feistel rounds and on (b) the fact that the first and last Feistel round produce collisions only with negligible probability. For (a) note that the [7] pseudorandom functions  $f : \{0, 1\}^m \rightarrow \{0, 1\}^{\ell(m)}$  obtain indistinguishability up to negligible factors not only for  $m = \text{poly}(n)$ , but whenever  $m \geq n + 1$  (In fact, the smaller the domain size is, the better the guaranteed security is). For (b) note that using pairwise independent functions for the first and last Feistel rounds produces collisions with probability  $\Theta(\sqrt{2^m})$  which is negligible as long as  $M = 2^m$  is. Therefore pseudorandomness is obtained for arbitrary proper  $M$ .

ever, will occur only with negligible probability, which suffices for our purpose of achieving computational indistinguishability.

The simpler construction relies on the fact that the [14] implementation of pseudorandom involutions directly applies the Luby-Rackoff construction of pseudorandom permutations (that are indistinguishable from the uniform distribution over all permutations  $\pi : [M] \rightarrow [M]$ ). Thus, it suffices to provide pseudorandom *permutations* for arbitrary proper domains. This is achieved via ‘cycle-walking’ (due to [3] and [10]) as follows. We consider an auxiliary pseudorandom permutation  $\pi' : [M'] \rightarrow [M']$  where  $M' = 2^{\lceil \log_2 M \rceil}$  is a power of 2. We define  $\pi(x)$  for  $x \in [M]$  by iterating  $\pi'$  until receiving an element in  $[M]$ . Namely,  $\pi(x) = \pi'^t(x)$  where  $t$  is the minimal index s.t.  $\pi'^t(x) \in [M]$ . To evaluate  $\pi(x)$  efficiently, we iteratively compute  $\pi'^j(x)$  until either  $\pi'^j(x) \in [M]$  or  $j = n$ , where the second case means failure to compute  $\pi(x)$  (the specific choice of  $n$  instead of any other  $n' = n^{\Theta(1)}$  is arbitrary). Ignoring the failures, if  $\pi'$  is uniformly random (w.r.t. the domain  $[M']$ ) then so is  $\pi$  (w.r.t. the domain  $[M]$ ). Therefore the pseudorandomness of  $\pi'$  implies the pseudorandomness of  $\pi$ . Finally, since  $M' \leq 2M$ , failures are detected only with negligible probability for a truly random  $\pi'$ , and therefore detected with negligible probability for a pseudorandom  $\pi'$  as well. The desired pseudorandomness follows. ■

## 6.4 Closure of $\sim$ Under Arithmetic Operations (proving Fact 2)

We gradually establish the retainment of  $\sim$  under various arithmetic operation, where  $F(1 \pm \delta)$  stands for some term  $E$  s.t.  $F(1 - \delta) \leq E \leq F(1 + \delta)$ .

1. Summation.  $C_i = \bar{C}_i(1 \pm \epsilon)$  clearly implies  $\sum_i C_i = (1 \pm \epsilon) \sum_i \bar{C}_i$ .
2. Multiplication.  $C = \bar{C}(1 \pm \epsilon)$ ,  $D = \bar{D}(1 \pm \epsilon)$  implies  $CD = \bar{C}\bar{D}(1 \pm \epsilon)(1 \pm \epsilon) = \bar{C}\bar{D}(1 \pm 3\epsilon)$  (the final inequality holds for all  $\epsilon \leq 1$ ).
3. Inverse. For any  $x$ ,  $(1 + x)^{-1} = 1 - x(1 - \frac{x}{1+x})$ . Whenever  $|x| \leq 0.5$ ,  $|\frac{1}{1+x}| \leq 2$  so  $1 - \frac{x}{1+x} = 1 \pm 2x = \pm 2$ . Thus,  $(1 + x)^{-1} = 1 \pm 2x$ .
4. Applying the multiplication estimate (with  $|\epsilon| \leq 1$ ) and then the summation estimate gives,  $\sum_{i=1}^{m_i} A_i^1 A_i^2 = (1 \pm 3\epsilon) \sum_{i=1}^{m_i} \bar{A}_i^1 \bar{A}_i^2$ . Similarly  $\sum_{i=1}^{m_i} B_i^1 B_i^2 = (1 \pm 3\epsilon) \sum_{i=1}^{m_i} \bar{B}_i^1 \bar{B}_i^2$ .
5. Applying item 4 gives

$$\left( \sum_{i=1}^{m_i} A_i^1 A_i^2 \right) / \left( \sum_{j=1}^{m_j} B_j^1 B_j^2 \right) = \left( (1 \pm 3\epsilon) \sum_{i=1}^{m_i} \bar{A}_i^1 \bar{A}_i^2 \right) / \left( (1 \pm 3\epsilon) \sum_{j=1}^{m_j} \bar{B}_j^1 \bar{B}_j^2 \right).$$

Finally, by item 3,  $(1 \pm 3\epsilon)^{-1} = 1 \pm 6\epsilon$  whenever  $\epsilon \leq 1/6$ , so  $(1 \pm 3\epsilon)/(1 \pm 3\epsilon)^{-1} = (1 \pm 3\epsilon)(1 \pm 6\epsilon) = 1 \pm 27\epsilon$ . ■

## 6.5 Various Calculations

**Fact 4** For  $N - (q - 1)k \leq N' \leq N$  and  $k, q \leq \text{poly}(\lg N)$  it holds that

$$\binom{N'}{k-1} p^{k-1} (1-p)^{N'-(k-1)} = \frac{N^{k-1}}{(k-1)!} p^{k-1} (1-p)^N \left( 1 \pm \Theta \left( \frac{qk^2}{N} + pqk \right) \right).$$

**Proof.** The fact follows immediately from the following:

1.  $\binom{N'}{k-1} / (k-1)! = \prod_{i=0}^{k-2} (N' - i)$ .
2.  $N^{k-1} > \prod_{i=0}^{k-2} (N' - i) > (N - (q-1)k - k)^{k-1} = N^{k-1} \left( 1 - \frac{qk}{N} \right)^{k-1} > N^{k-1} \left( 1 - \Theta \left( \frac{qk^2}{N} \right) \right)$ .
3.  $1 < (1-p)^{N'-(k-1)} / (1-p)^N < (1-p)^{-qk} = 1 + \Theta(pqk)$ .

(the two  $\Theta$  estimates are implied by  $(1+x)^r = 1 \pm \Theta(xr)$  valid whenever  $|x|, |xr| = o(1)$  (regardless of whether  $x, r \geq 0$ )). ■

**Fact 5** Whenever  $p, \ell/N$  are negligible it holds that

$$\ell \binom{N-1}{\ell} p^\ell (1-p)^{N-1-\ell} \sim \binom{N-1}{\ell-1} p^{\ell-1} (1-p)^{N-1-(\ell-1)} p N.$$

**Proof.** The fact follows immediately from  $1-p \sim 1$  and the following estimations.

$$\begin{aligned} \ell \binom{N-1}{\ell} &= \frac{\ell}{\ell!} \prod_{i=0}^{\ell-1} (N-i) \\ &= (N-\ell+1) \left[ \prod_{i=0}^{(\ell-1)-1} (N-i) / (\ell-1)! \right] \\ &\sim N \left[ \prod_{i=0}^{(\ell-1)-1} (N-i) / (\ell-1)! \right] \\ &= N \binom{N-1}{\ell-1}. \quad \blacksquare \end{aligned}$$

**Fact 6** Let  $\kappa = \prod_{i=1}^m (S+1-2i)^{-1}$ , with  $m < qd_{\max} = q \lg N \lceil pN \rceil$  and  $S = pN(N-1)(1 \pm \epsilon)$  for  $\epsilon = \frac{\lg \lg N}{\sqrt{pN}}$ . Then  $\kappa = [pN(N-1)]^{-m} (1 \pm \delta)$  for some negligible  $\delta$  which is independent of the specific value of  $S$  and  $m$ .

**Proof.** We show that (i)  $\kappa = S^{-m} (1 \pm \delta_1)$  and that (ii)  $S^{-m} = (pN(N-1))^{-m} (1 \pm \delta_2)$ , for negligible  $\delta_1, \delta_2$  which are independent of the specific value of  $S$  and  $m$ . As  $(1 \pm \delta_1)(1 \pm \delta_2) = 1 \pm 3(\delta_1 + \delta_2)$  for any  $\delta_1, \delta_2 < 1$ , the claim follows.

For (i) it clearly holds that,

$$S^{-m} < \kappa < (S-2m)^{-m} = S^{-m} \left( 1 - \frac{2m}{S} \right)^{-m} < S^{-m} \left( 1 + \frac{4m^2}{S} \right),$$

where the final  $<$  is given by  $(1 - x)^{-\ell} < 1 + 2\ell x$  valid for all  $x, \ell x < 1/4$ . This condition on  $x, \ell$  holds since  $m^2 = o(S)$  (as discussed bellow). Now, as  $S \sim pN(N - 1)(1 \pm \epsilon)$  then  $\frac{4m^2}{S} < \frac{4d_{\max}^2}{S} \sim \frac{4(q \lg N \lceil p(N-1) \rceil)^2}{pN(N-1)}$ . Our original  $\frac{(\lg N)^{O(1)}}{N}$  upper bound on  $p$  implies that the nominator is  $poly(\lg N)$  bounded, while our original  $\frac{(\lg N)^{\omega(1)}}{N^2}$  lower-bound on  $p$  implies that the denominator is  $(\lg N)^{\omega(1)}$ . Thus  $\frac{4m^2}{S}$  can be bounded by some negligible  $\delta_1$  which is independent of the specific value of  $S$  and  $m$ . Part i) follows.

For (ii) Note that  $S^{-m} = (pN(N - 1))^{-m} (1 \pm \epsilon)^{-m}$ . We use the fact  $(1 \pm \epsilon)^{-m} = (1 \pm 2\epsilon m)$  valid for any  $\epsilon \cdot m = o(1)$ . This condition on  $\epsilon \cdot m$  holds as follows. First,  $\frac{(\lg N)^{\omega(1)}}{N^2} \leq p$  implies that  $(\sqrt{p}N)^{-1}$  is negligible. Next,  $p \leq \frac{poly(\lg N)}{N}$  implies that  $\lceil pN \rceil \leq poly(\lg N)$ . Therefore,  $\epsilon \cdot m < \frac{\lg \lg N}{\sqrt{p}N} \cdot q \lg N \lceil pN \rceil = (q \lg N \lg \lg N) \frac{\lceil pN \rceil}{\sqrt{p}N}$  is negligible, and so is  $2\epsilon m$ . Part ii) follows. ■