The Theory of Zeta Graphs with an Application to Random Networks

Christopher Ré
Stanford University
chrismre@cs.stanford.edu

ABSTRACT

Social, biological, and cyberphysical networks generate some of the most intriguing and valuable sources of data on the planet. For at least the last two decades, researchers have attempted to create formal (typically stochastic) models of these networks. We examine the database theory questions raised by these new models. We study a simple extension of Erdös–Rényi models that we call Zeta graphs. Zeta graphs are related to multiple-valued zeta functions, and we show that the expectation of a conjunctive query can be written as a linear combination of multiple-valued zeta functions. For queries on graphs, we use our results to devise a complete decision procedure for whether the probability that a query is true tends to 1 as the domain size tends to infinity. We apply our theory of Zeta graphs to describe the set of conjunctive graph queries that are true with probability 1 in another graph model in the literature that was described by Callaway, Hopcroft, Kleinberg, Newman, and Strogatz.

Categories and Subject Descriptors

H.2.m [**Database Applications**]: Miscellaneous

General Terms

Database Theory, Random Graphs

Keywords

Database Theory, Multiple-valued Zeta Function, Erdös-Rényi Graphs, Preferential Attachment

1. INTRODUCTION

The scientific and commercial value of cyberphysical, social, and biological networks is exploding. Motivated by this explosion, models for such networks have been a hot topic over the last two decades [1, 2, 6, 17, 19, 24]. As with any formal model, we hope to find a mathematically elegant, tractable model that captures the salient features of real-world graphs. In turn, the model will ideally allow one to derive properties of the graph that yield new insight into the underlying structure of real-world networks. Newman et al. state that exactly solvable models have already lead to a wealth of insights and are a goal of this line of research [20]. Our goal is to apply ideas from database theory to contribute to this exciting line of work.

Unfortunately, there is no single agreed-on graph model to drill into, as shown by the plethora of models for such networks [2, 6, 17]. Each model seems to capture some—but not all—aspects of real network graphs [21]. We do not intend to add another voice to this debate; however, we do intend to understand to what extent these models may provide new questions and opportunities for database theoreticians. In particular, the database community has used random models of data to explore a wide variety of questions, e.g., privacy [10, 13], and fundamental questions about the powers of logic [18, 19] including the celebrated zero-one law for first order logic [12, 14].

One popular model that motivated our work was devised by Callaway, Hopcroft, Kleinberg, Newman, and Strogatz (henceforth, CHKNS) and is described by the following procedure [6, p. 1]:

> *At each time step, a new vertex is added. Then, with probability δ, two vertices are chosen uniformly at random and joined by an undirected edge.*

CHKNS's model is intriguing: it has an elegant generative description and captures some of the structure of real networks. While this model is elegant, the generative description makes it difficult to perform the computations that are needed to prove logic-based database-style theorems [10, 18, 19]. As an application of our main results, we are able to completely answer questions about the logical Theory of these graphs,[1] i.e., we

[1] Following Enderton [11, p. 155] and Libkin [18, p. 241],

10.5441/002/icdt.2014.21

describe the set of queries such that each query is true with probability 1 as the domain size tends to infinity in CHKNS's model (and its directed-graph analog). This result allows us to understand a fundamental question: *Which graph structures appear in this model, and which do not?* This is a component of the central theoretical question in network motif theory [21, 23]: *Is a particular pattern statistically significant, or should one expect such a structure due only to random chance?* Our work provides a start on this question using the tools of database theory.

We focus on a class of graphs called *Zeta graphs* that have a simple declarative description and capture a key technical aspect of network models including CHKNS and the famous Albert and Barabási preferential attachment model [2, p. 73]. Formally, Zeta graphs are easy to define:

Definition 1.1. *For each $N \geq 1$, let $\mathcal{Z}_N$ be a probability distribution on graphs in which the nodes are the integers $[N] = \{1, \ldots, N\}$ and the probability of each edge $(i, j) \in [N]^2$ is an independent random variable given by the following expression:*

$$\Pr_{\mathcal{Z}_N}[(i,j) \in E] = u^{-1} \text{ where } u = \max\{i, j\}$$

Intuitively, this model is close to an Erdös–Rényi (ER) model , since each edge is assigned an independent probability. In contrast to traditional Erdös–Rényi graphs, in which the probability of all edges is the same, the probability of an edge depends on its "latest arriving" endpoint. This captures a property of many network models (including CHKNS's): that nodes enter the graph at different times, and that those nodes that enter later have a different probability of making connections than those that have been around for a long time. This late-arrival property is shared by many models including CHKNS and preferential attachment models. Of course, higher fidelity network models often take other factors into account to more finely model the network, e.g., degree distributions [2] or hyperbolicity [7]. In this work, however, we focus on this one critical aspect as it already poses non-trivial technical challenges over traditional ER models.

New Technical Results. The goal of the first part of our work is to develop techniques to compute exact and asymptotic approximations for the probability of conjunctive queries on Zeta graphs. Using standard techniques for random graphs (Janson's inequality), the probability of a Boolean query q can be related to the expected number of tuples returned by a second query Q; Q has the same body as q, but Q's head contains all variables in the body of q. Our first technical result is an algorithm that expresses the expected number of tuples of an arbitrary conjunctive query (with ordered constraints) as a linear combination of a family of special functions, called the *multiple-valued zeta functions* (MVZs).[2] MVZs are a generalization of Riemann's famous zeta function $\zeta(s) = \sum_{n=1}^{\infty} n^{-s}$ to multiple arguments or values, e.g., for $\zeta(s_1, s_2) = \sum_{0<n_1<n_2}^{\infty} n_1^{-s_1} n_2^{-s_2}$. MVZs have been studied for the last few decades, and they arise in multiple areas (e.g., quantum field theory [4] and analytic number theory [16]). For queries with totally ordered variables, this algorithm is efficient (linear time in the size of the query). For more general queries, our algorithm takes exponential time, which is likely unavoidable as we show that finding the expectation is ♯P-Hard.

It is not clear that MVZs are substantially easier to deal with than their original formulation as conjunctive queries. However, MVZs are well studied. In particular, there are rapidly converging approximations for MVZs, which enables efficient computation [9]; there is a recent theory of asymptotic approximation [8], and there is a rich theory of combinatorial and algebraic identities [24].[3] By mapping our problem to MVZs, we gain a host of computational and approximation results from the last 20 years.

We then turn to the problem of computing the asymptotic probability of a family of CQs on Zeta graphs as the size of the domain tends to infinity. For graph-queries (i.e., each relation is binary), we are able to give a precise characterization of the set of conjunctive queries that have probability 1. This result is inspired by the fundamental results about other random graph models [12, 19, 22]. We use this model to analyze the CHKNS model, which was previously proposed.

Application: Back to CHKNS's Model. Our goal is to understand the theory of the CHKNS model. Roughly, there are two main technical challenges:

1. CHKNS has edge probabilities that are close to our Zeta model, but they are not exactly the same. As a result, we need to keep track of the approximation quality with some care as $N \to \infty$.
2. CHKNS's model introduces correlations between edge random variables, which are not present in Zeta graphs. We need to develop bounds to compensate for these correlations.

We begin with some technical work on CHKNS's graphs. In the case of $\delta = 1$, we show that, in CHKNS's model, the probability that a disjunction of edge random variables can be computed in closed form (in terms of Euler's Gamma function). Our technique uses a well-known correspondence between the Gamma function and products of terms that are polynomial functions of the index. Using this relationship, we can compute any propositional formula using the inclusion–exclusion formula, in which the number of terms is bounded by a function of the query size.

we call this set the Theory and use the notation Th. We formally define this concept in Section 2.

[2] This connection is how we came to the name *Zeta graphs*.

[3] There are some minor technical issues that prevent us from directly reusing these results, but we are able to use the main techniques to derive the tools that we need: Euler–Maclaurin summations [8] or see the textbook [15, p. 469].

In traditional Erdös–Rényi random graphs, one has a simple closed formula to compute the probability of either the conjunction or the disjunction of a set of events; in our situation, computing the probability of such conjunctions and disjunctions seem to require using an inclusion–exclusion formulae. To overcome this obstacle, we use a standard technique: we develop a pair of models, an upper- and lower-bound model for CHKNS's model, in the sense that the probability of any monotone formula is only higher (resp. lower) in the upper (resp. lower) approximation. These approximations have two key additional properties: (1) their edge probabilities are independent and are essentially appropriately scaled Zeta graphs, and (2) for any propositional formula of positive edge atoms ϕ, there is a constant $\gamma > 0$ (see Proposition 4.3) such that the probability $\Pr[\phi]$ is no more than a factor γ smaller on the independent model than on CHKNS. We then show that the Theory of both the upper- and lower-bound models coincide. Since the Theory of CHKNS is sandwiched in between these two models, this implies that the Theory of CHKNS coincides as well. Using these ideas, we are able to describe the Theory of conjunctive queries on CHKNS's graphs and show that CHKNS does not have a zero-one law.

Outline. In Section 2, we describe CHKNS's model, our proposed model described above, the relevant portion of the algebraic theory from MVZs, and some standard techniques for this area, e.g., Janson's inequality. In Section 3, we describe our techniques to compute the probability of conjunctive queries on Zeta graphs. In Section 4, we describe our main application to the CHKNS model. We describe related work in Section 5.

2. PRELIMINARIES

We begin by introducing some notation, defining the queries we will consider, and some background on MVZs.

2.1 Queries and Probabilistic Databases

We assume that there exists an infinite set of relational symbols, and define CQ to be the set of Boolean conjunctive queries that do not contain constants, but may contain comparisons (order predicates) among the variables.[4] Denote by $\mathbf{var}(q)$ the set of variables in a query q. Let CQ^{TO} be the subset of CQ such that for any q the variables $\mathbf{var}(q)$ are totally ordered by constraints. That is, for any two distinct variables x_i and x_j in a query $q \in \mathrm{CQ}^{TO}$, we can deduce that $x_i < x_j$ or *vice versa.* $R(x,y), R(y,z), x < y, y < z$ is totally ordered (we can deduce $x < z$ transitively), while $R(x,y), R(y,z), x < z$ is not (as we cannot deduce whether $y < x$ or $x < y$.) We refer to this total order as the total order on variables of q.

An important special case is when a query contains only a single binary relational symbol (R): we denote this class of queries as CQ_{1B}. Since queries in CQ_{1B} can be thought of as graphs, we use ideas from graph theory, e.g., cycles, paths, and connected components. We use CQ_{1B}^{TO} to denote such queries in which all variables are totally ordered.

As we will see below, the probability that a Boolean query q is satisfied is related to the expectation of a related query that we denote by Q. Q can be constructed from q by first copying the body of q and then adding all variables in the body of q to the head of Q (we call Q the *full query corresponding* to q). We illustrate such a pair of queries (q_1, Q_1); here q_1 checks for a path of length 2:

$$\begin{aligned} q_1() &= R(x,y), R(y,z), x < y, y < z \\ Q_1(x,y,z) &= R(x,y), R(y,z), x < y, y < z \end{aligned}$$

That is, Q_1 adds all variables in the body of q_1 to the head of Q_1. Our notation is that lowercase symbols, e.g., q_1 above, always denote Boolean queries, while the same query with a capitalized symbol, e.g., Q_1, will always be the corresponding full query. Queries denoted with uppercase letters always denote full queries.

Databases. We consider probabilistic databases $(\mathcal{I}, \mu)$ that consist of a finite set of *possible worlds* $\mathcal{I}$ and a corresponding probability measure μ:

$$\mu : \mathcal{I} \to [0,1] \text{ such that } \sum_{I \in \mathcal{I}} \mu(I) = 1$$

Typically, $\mathcal{I}$ will be huge (e.g., all subsets of a given set of tuples) and so the measure μ will often be given implicitly. For example, we often consider tuple-independent databases in which μ is given by a product. In this case, we are given a set of tuples T and a function $p : T \to [0,1]$ (the marginal function) such that

$$\mu(I) = \prod_{t:t\in I} p(t) \times \prod_{t:t\notin I} (1 - p(t))$$

Query Answering. Given a probabilistic database $(\mathcal{I}, \mu)$, we talk about the probability that a Boolean query q (in any language) is true or satisfied by regarding q as a function from $\mathcal{I} \to \{0,1\}$ and we define

$$\Pr_{(\mathcal{I},\mu)}[q] = \sum_{I \in \mathcal{I}: q(I)=1} \mu(I)$$

Families of Databases. Our technical results are concerned with countably infinite families of databases,

$$\{(\mathcal{I}_1, \mu_1), (\mathcal{I}_2, \mu_2), \dots\}.$$

We will take limits in these sets. For this work, a particularly important set of families is called Zeta databases. We first introduce Zeta databases with a single relational symbol, and then we extend to multiple relational symbols (stipulating independence between events in each table).

Given a relational symbol $R(x_1, \dots, x_k)$ of arity k, for each $N \geq 1$, we define the N^{th} element of this family to

[4] Many of our results can handle queries with constants in a straightforward way, e.g., the tools that allow us to compute probabilities. However, our results about which queries are satisfied with probability 1 in a model will likely have different characterizations.

be the pair $(\mathcal{Z}_N^R, \mu_N^R)$, where $\mathcal{Z}_N^R$ is all instances of R on the $[N]^k$, i.e., all subsets of $\{R(\bar{c}) \mid \bar{c} \in [N]^k\}$ and μ_N^R is defined as:

$$\mu_N^R(I) = \prod_{\bar{c}\in[N]^k} \left(u(\bar{c})^{-1}\mathbf{1}_{[R(\bar{c})\in I]} + (1 - u(\bar{c})^{-1})\mathbf{1}_{[R(\bar{c})\notin I]}\right)$$

where $u(\bar{c}) = \max_{i=1}^k c_i$ and $\mathbf{1}_{[R(\bar{c})\in I]}$ (resp. $\mathbf{1}_{[R(\bar{c})\notin I]}$) is the indicator function for $R(\bar{c}) \in I$ (resp. $R(\bar{c}) \notin I$). When $k = 2$, this matches the description given in Definition 1.1.

Given a schema $\sigma = \{R_1, \ldots, R_m\}$ with multiple relational symbols and $N \geq 1$, we extend this definition by insisting that the events are independent. We denote the N^{th} member of the resulting Zeta database as $\mathcal{Z}_N^\sigma = \times_{i=1}^m \mathcal{Z}_N^{R_i}$ and $\mu_N^\sigma(I) = \prod_{i=1} \mu_N^{R_i}(I_{|R_i})$ where $I_{|R_i}$ denotes the instance restricted to (or projected onto) the relation R_i.

Given a query q, we will often assume implicitly that we are considering a database that contains only the schema symbols mentioned in q. As a result, we write $\Pr_{\mathcal{Z}_N}[q]$ as a shorthand for the probability that the query is true on the N^{th} element of the Zeta family that has the appropriate schema from q instead of the heavier, but less ambiguous, notation $\Pr_{(\mathcal{Z}_N^{\sigma(q)}, \mu_N^{\sigma(q)})}[q]$.

Theory. Given a class of queries $\mathcal{Q}$ (e.g., $\mathcal{Q}$ = CQ) and a countably infinite family of instances $\mathcal{I}_1, \mathcal{I}_2, \ldots$, our technical results deal with the limits as N goes to infinity of the quantity $\Pr_{\mathcal{I}_N}[q]$. Given an infinite family $\mathcal{I}_1, \mathcal{I}_2, \ldots$ we write $q \in \text{Th}(\mathcal{I}_\infty, \mathcal{Q})$ whenever

$$\lim_{N\to\infty} \Pr_{\mathcal{I}_N}[q] = 1$$

We are especially interested in $\text{Th}(\mathcal{Z}_\infty, \text{CQ}_{1B})$, as CQ_{1B} captures graph patterns, which are of independent interest and form the core of our later technical results.

2.2 Probabilistic Tools

One tool that we use is Janson's inequality, which has also been used extensively in random graph work (see Alon and Spencer [3]). The typical usage is to relate the probability that a Boolean query is satisfied to the expectation of a corresponding full query. We restate Janson's inequality in the terminology of this paper and an easily derived corollary.

LEMMA 2.1 (JANSON'S INEQUALITY). *Given a Boolean query q with variables $\bar{x}$, let Q denote its corresponding full query. Define λ and Δ:*

$$\begin{aligned} \lambda = & \ \mathbb{E}[Q(\bar{x})] = \textstyle\sum_{\bar{c}\in[N]^{|x|}} \mathbb{E}[Q(\bar{c})] \ \textit{and} \\ \Delta = & \ \textstyle\sum_{\bar{c},\bar{d}\in[N]^k : Q(\bar{c}) \textit{ properly overlaps } Q(\bar{d})} \mathbb{E}[Q(\bar{c}) \wedge Q(\bar{d})] \end{aligned}$$

where $Q(\bar{c})$ properly overlaps $Q(\bar{d})$ if $Q(\bar{c}) \neq Q(\bar{d})$, and there is some subgoal that is identical in each $Q(\bar{c})$ and $Q(\bar{d})$. Then,

$$1 - \Pr[q] \leq \exp\{-\lambda + \Delta/2\}$$

and if $\Delta > \lambda$ then

$$1 - \Pr[q] \leq \exp\{-\lambda^2/(\lambda + \Delta)\}$$

A corollary that is often easier to apply for the results of this paper is the following:

COROLLARY 2.1. *With the notation of Lemma 2.1, for $N \geq 0$, if $\mathbb{E}_{\mathcal{I}_N}[\lambda] = \omega(1)$ and $\mathbb{E}_{\mathcal{I}_N}[\Delta] = o(\lambda^2)$, then*

$$\lim_{N\to\infty} \Pr_{\mathcal{I}_N}[q] = 1.$$

where all asymptotic notation is with respect to N.

This corollary is a just restatement of the second case of Janson's inequality: the conditions of the Corollary 2.1 imply that $\lambda + \Delta = o(\lambda^2)$ and so $\exp\{-\lambda^2/(\lambda + \Delta)\} = o(1)$.

2.3 Zeta Functions

To apply Janson's inequality, we need to compute the expected number of tuples returned by a query on the probability model. In traditional Erdös–Rényi graphs, this computation is typically straightforward since the probability of every edge is the same (see Dalvi et al. [10]). However, for $\mathcal{Z}_N$ the probability of an edge varies and, as we describe below, the computation is related to ζ functions. We introduce ζ functions and their relevant properties.

Arguably the most famous of these functions is Riemann's Zeta function, which is denoted ζ and is defined as:

$$\zeta(s) = \sum_{i=1}^{\infty} i^{-s}$$

This function, ζ, is one of the most famous in mathematics; the location of the zeros of this function in the complex plane is the subject of the Riemann hypothesis.

We consider a classical generalization of this function [24], called MVZs. We will also consider a finite variant. In particular, for k numbers $s_1, \ldots, s_k$ and $N \geq k$ we define:

$$\zeta^N(s_1, s_2, \ldots, s_k) = \sum_{0<x_1<x_2<\cdots<x_k\leq N} \prod_{j=1}^{k} x_j^{-s_j}$$

Clearly, $\lim_{N\to\infty} \zeta^N(s_1) = \zeta(s_1)$.[5] We know that $\zeta^N(1)$ is the N^{th} harmonic number.

Zeta Notation. Following the survey by Zudlin [24] and standard practice for MVZs, we will often denote the arguments to MVZs using the following compressed notation: if a value appears (consecutively) several times, then we enclose the value in set braces and subscript this term with the number of times that this value occurs. For example, we shall abbreviate $\zeta^N(0, 0, 1, 2, 2, 2)$ as $\zeta^N(\{0\}_2, 1, \{2\}_3)$.

3. ZETA DATABASES

The main result of this section is a simple-to-check characterization of $\text{Th}(\mathcal{Z}_\infty, \text{CQ}_{1B})$. Given q and its corresponding full query Q, we begin with an exact algorithm to compute $\mathbb{E}_{\mathcal{Z}_\infty}[Q]$ when $q \in CQ$, and describe the hardness of computing this quantity. We then describe our main result.

[5] This is true definitionally, but the limit may be divergent.

3.1 Exact Algorithms in terms of ζ

Given (q, Q), our goal in this section is to compute the expected number of tuples returned by a query Q for some $q \in$ CQ on $\mathcal{Z}_N$, denoted $\mathbb{E}_{\mathcal{Z}_N}[Q]$. We first state a connection between $\mathbb{E}_{\mathcal{Z}_N}[Q]$ when q is in the class CQ^{TO}.

3.1.1 Exact Expectations for CQ^{TO} in Terms of ζ^N

We fix some positive N (to avoid triviality, assume N is larger than the number of variables). We illustrate the connection between MVZs and the expected number of tuples of a query Q with an example:

Example 3.1. *Consider the query q_2 defined as*

$$q_2 = R(x, y), R(y, z), x < y, y < z. \textit{Then,}$$

$$\mathbb{E}_{\mathcal{Z}_N}[Q_2(x, y, z)] = \sum_{1 \leq i_1 < i_2 < i_3 \leq N} i_2^{-1} i_3^{-1} = \zeta^N(0, 1, 1)$$

A quick calculation shows that we can resolve this even further.

$$\begin{aligned} \zeta^N(0,1,1) \quad &= \textstyle\sum_{1 \leq i_2 < i_3 \leq N} (i_2 - 1) i_2^{-1} i_3^{-1} \\ &= \textstyle\sum_{1 \leq i_2 < i_3 \leq N} i_3^{-1} - \zeta^N(1,1) \\ &= \textstyle\sum_{1 \leq i_3 \leq N} (i_3 - 1) i_3^{-1} - \zeta^N(1,1) \\ &= \zeta^N(0) - \zeta^N(1,1) - \zeta^N(1) \\ &= N - \zeta^N(1,1) - H_N \end{aligned}$$

where H_N is the N^{th} harmonic number. Using distinct relational symbols does not change the expectation, e.g., $Q_3(x, y, z) = R(x, y), S(y, z), x < y < z$, then $\mathbb{E}_{\mathcal{Z}_\infty}[Q_3] = \mathbb{E}_{\mathcal{Z}_\infty}[Q_2]$. However, we do need to be careful about the ordering of the variables. Consider the query,

$$Q_{2'}(x, y, z) = R(x, y), R(y, z), x < z < y.$$

It yields a different function that we can also resolve:

$$\begin{aligned} \mathbb{E}_{\mathcal{Z}_\infty}[Q_{2'}(x,y,z)] = \quad & \textstyle\sum_{1 \leq i_1 < i_2 < i_3 \leq n} i_3^{-2} \quad && = \zeta^N(0,0,2) \\ = \quad & \textstyle\sum_{i_3} \binom{i_3 - 1}{2} i_3^{-2} && \sim \tfrac{1}{2} N \end{aligned}$$

The change in variable order cut the expectation in half. Later (Example 3.3), we give an example in which the difference in expectation between two orderings is arbitrarily large (i.e., the difference depends on N). □

To generalize this example, we define some notation. Given a query $q \in \mathrm{CQ}^{TO}$ with subgoals $g_1, \ldots, g_t$, we say that a subgoal is redundant if $g_i = g_j$. We first remove all redundant subgoals, which can be done in linear time. Let the variables of q be $x_1, \ldots, x_k$, and without loss of generality, assume they are in order (i.e., that $x_i < x_j$ in the total order specified by q.). Define $s_i(q)$ as follows for $i = 1, \ldots, k$

$$s_i(q) = |\{j \in [t] \mid i = \max \mathbf{var}(g_j)\}|$$

Proposition 3.1. *Using the notation above, given $q \in \mathrm{CQ}^{TO}$ with k variables, $x_1, \ldots, x_k$, and Q its corresponding full query, let $s_i = s_i(q)$. Then,*

$$\mathbb{E}_{\mathcal{Z}_N}[Q(x_1, \ldots, x_k)] = \zeta^N(s_1, \ldots, s_k). \qquad (1)$$

Proof. Since $q \in \mathrm{CQ}^{TO}$, any homomorphism of q must be injective (since the images of all of the variables of q must be totally ordered, and so no two variables can be equated). Hence, the result of $Q(\bar{x})$ can be written as

$$Q(\bar{x}) = \cup_{\bar{c}: c_1 < \cdots < c_n} Q(\bar{c}).$$

Since the union is disjoint,

$$\mathbb{E}_{\mathcal{Z}_N}[Q] = \sum_{\bar{c}: c_1 < \cdots < c_n} \mathbb{E}_{\mathcal{Z}_N}[Q(\bar{c})].$$

For any such query, since there are no redundant subgoals all edges must be present. Then, $\mathbb{E}_{\mathcal{Z}_\infty}[Q(\bar{c})] = \prod_{j=1}^{t} \Pr[g_j] = \prod_{j=1}^{t} u_j^{-1}$ where $u_j = \max\{c : c \in g_j\}$. In turn, one can rewrite this as $\prod_{i=1}^{k} c_i^{-s_i}$, grouping by the variables. Summing over all such $\bar{c}$ is the claimed MVZ, $\zeta^N(s_1, \ldots, s_k)$. □

3.1.2 Extending to All CQs

We reduce the problem of computing $\mathbb{E}_{\mathcal{Z}_\infty}[Q]$ for $q \in$ CQ to the problem of computing the expectation for several queries in CQ^{TO}, for which the algorithms of the last section apply. A straightforward way to do this is to essentially consider all possible orderings and unifications that are consistent with the comparisons in q. We describe this idea more precisely below, and we give an algorithmic description so that we can discuss the running time of our procedure.

Given a query $q \in$ CQ, we perform a two-step process to describe the set of all mappings of q to CQ^{TO}, denoted $\mathrm{Hom}(q, \mathrm{CQ}^{TO})$:[6] (1) We define the set of all mappings of variables to variables that respect the comparisons, and (2) we construct canonical images under each mapping. In step (1), for a query q denote by $H_k(q)$ the set of surjective maps $\sigma : \mathbf{var}(q) \to [k]$ that are also order-preserving maps, i.e., if $x_i < x_j$ then $\sigma(x_i) < \sigma(x_j)$. Finally, define $\mathrm{Hom}(q, \mathrm{CQ}^{TO}) = \cup_k H_k$, the set of all such mappings from q.[7]

Each $\sigma \in H_k(q)$ can be associated with a query, $q^{+\sigma}$, which is its homomorphic image, as follows: $q^{+\sigma}$ is initialized to a copy of q. Then, for each pair $(x_i, x_j) \in \mathbf{var}(q)^2$, if $\sigma(x_i) = \sigma(x_j)$ then we add the equality $x_i = x_j$ to $q^{+\sigma}$, otherwise without loss $\sigma(x_i) < \sigma(x_j)$ and we add $x_i < x_j$ to $q^{+\sigma}$ (if it is not already present). Notice that $q^{+\sigma}$ is in CQ^{TO} and all such homomorphic images arise this way, justifying the $\mathrm{Hom}(q, \mathrm{CQ}^{TO})$ notation above.

Example 3.2. *Let $q = R(x_1, x_2), R(x_2, x_3), x_1 < x_3$. We denote the mapping σ with its domain a triple subscript: $\sigma_{y_1 y_2 y_3}$ denotes the mapping that $\sigma(x_i) = y_i$ for $i = 1, 2, 3$. Then we have $H_3(q) = \{\sigma_{213}, \sigma_{123}, \sigma_{132}\}$, $H_2(q) = \{\sigma_{112}, \sigma_{122}\}$, and $H_1(q) = \emptyset$. The second stage results in five queries (one for each mapping above),*

[6] Observe that the set of homomorphisms from q is in a bijection with homomorphism from Q (the bijection is given by syntactically identical homomorphisms). Thus, we abuse notation and think about homomorphisms as being from either structure into CQ^{TO}.

[7] The support of $\cup_k H_k$ is finite, since $H_j = \emptyset$ for all $j \geq |\mathbf{var}(q)|$.

e.g.,

$$\begin{aligned} Q^{+\sigma_{213}}(x_1, x_2, x_3) &= R(x_1, x_2), R(x_2, x_3), x_2 < x_1 < x_3 \\ Q^{+\sigma_{112}}(x_1, x_1, x_2) &= R(x_1, x_1), R(x_1, x_2), x_1 < x_2 \end{aligned}$$

With this, we can prove the main result for CQs:

Proposition 3.2. *With the notation above:*

$$\mathbb{E}_{\mathcal{Z}_N}[Q] = \sum_{\sigma \in \text{Hom}(Q, \text{CQ}^{TO})} \mathbb{E}_{\mathcal{Z}_N}[Q^{+\sigma}]$$

Proof. First, on any database instance $Q(I)$ is contained in $\bigcup_{\sigma \in \text{Hom}(Q,\text{CQ}^{TO})} Q^{+\sigma}(I)$, since any homomorphism from Q to I must be a mapping $\sigma \in \text{Hom}(Q, \text{CQ}^{TO})$. On the other hand, all $q^{+\sigma}$ are homomorphic images of Q, so $\bigcup_\sigma Q^{+\sigma} \subseteq q$. To ensure that the sum does not over count, it suffices to observe that the union is disjoint, since the comparisons force the output of $Q^{+\sigma}$ and $Q^{+\sigma'}$ to be disjoint if $\sigma \neq \sigma'$. □

Straightforwardly applying this proposition gives an exponential time algorithm (in the size of $|Q|$) for computing the expectation for any $Q \in \text{CQ}$. A natural question is: *can one create a substantially more efficient algorithm, e.g., in polynomial time?* We show next that the answer is likely no.

Proposition 3.3. *Given a* $q \in \text{CQ}$*, the problem of computing* $\mathbb{E}_{\mathcal{Z}_N}[Q]$ *is* ♯P*-Hard.*

Dalvi et al. [10] establish NP-hardness of the related problem over standard random graphs. In their reduction, the source of the difficulty is the unifications (e.g., the size of the automorphism group of q essentially determines the complexity of the algorithm). However, even for q with trivial automorphism groups (e.g., if any relational symbols appear in q's body at most once), the ♯P-hardness result still holds. The reason is that one can encode the problem of counting all extensions of a partial order to a total order (i.e., counting all linear extensions); a problem which admits a *fully polynomial time approximation scheme* (FPTAS). The standard FPTAS for counting linear extensions can be directly adapted to approximate $\mathbb{E}_{\mathcal{Z}_N}[Q]$ [5]. Thus, for $q \in \text{CQ}$, our simple algorithm may be close to optimal.

3.2 Application: $\text{Th}(\mathcal{Z}_\infty, \text{CQ}_{1B})$

We now describe the set of Boolean conjunctive queries in CQ_{1B} whose probability approaches 1 as $N \to \infty$.

3.2.1 Asymptotic Approximation of Expectation

For a query q, to compute $\lim_{N\to\infty} \Pr_{\mathcal{Z}_N}[q]$, our technique will be to use Janson's inequality (Lemma 2.1). To apply Janson's inequality, we need to estimate $\mathbb{E}_{\mathcal{Z}_N}[Q]$, in which Q is the full query corresponding to q.

Asymptotic Properties. We prove a useful proposition that roughly describes the rate of growth of ζ in terms of the partial sums of the arguments. This is a minor extension of Costerman et al. [8]:

Lemma 3.1. *Let* $\bar{s} \geq \bar{0}$*, let* $S_k(\bar{s}) = \sum_{i=k}^r (s_i - 1)$ *and define* $S_{\min} = \min_k S_k(\bar{s})$*. Then,*

$$\zeta^N(\bar{s}) = \begin{cases} \mathcal{O}(1) & \textit{if } S_{\min} > 0 \\ \Theta(\log^k N) & \textit{if } S_{\min} = 0 \textit{ and } k = |\{i \mid S_i = 0\}| \\ \tilde{\Theta}(N^{-S_{\min}}) & \textit{if } S_{\min} < 0 \end{cases}$$

Here $\tilde{\Theta}(\cdot)$ *hides factors up to* $\log^{|\bar{s}|} N$*.*

That is, $\zeta(\bar{s})$ is bounded if and only if $S_{\min} > 0$. For the lower bound, if any suffix $\bar{s}$ would result in a divergent sum, then the whole sum diverges. Second, on the upper-bound side, we can construct upper bounds using an integral to upper bound the sum [15, p. 469] or by using the Euler–Maclaurin summation formula.

The final ingredient in the proof of Lemma 3.1 is a majorization-type inequality: given two vectors $\bar{s}$ and $\bar{t}$ of the same arity, call it r, we write $\bar{s} \sqsubseteq \bar{t}$ if $\sum_{k=1}^r s_i \leq \sum_{k=1}^r t_j$ for each $k = 1, \ldots, r-1$. With this notation,

Proposition 3.4. *If* $\bar{s} \sqsubseteq \bar{t}$*, then* $\zeta(\bar{s}) \geq \zeta(\bar{t})$*.*

We highlight the result as it allows us to easily determine which ordering of the vertices of a query in CQ_{1B}^{TO} graph has the largest expectation.

3.2.2 The Theory of $\text{Th}(\mathcal{Z}_\infty, \text{CQ}_{1B})$

First, we begin with an example that shows that to compute the probability of a query, it is not enough to consider just the expectation; one needs also to consider the logical structure of the query.

Example 3.3 (K_4 with and without a tail).

$$\begin{aligned} q_{K_4} = \quad & \textstyle\bigwedge_{(i,j)\in\binom{[4]}{2}} R(x_i, x_j) \wedge \bigwedge_{k=1}^3 (x_k < x_{k+1}) \\ q_{K_4+T} = \quad & \textstyle\bigwedge_{(i,j)\in\binom{[4]}{2}} R(x_i, x_j) \wedge R(x_4, x_5) \\ & \textstyle\bigwedge_{k=1}^4 (x_k < x_{k+1}) \end{aligned}$$

From Lemma 3.1, one can see that $\mathbb{E}_{\mathcal{Z}_\infty}[Q_{K_4}] = \mathcal{O}(1)$ *and* $\mathbb{E}_{\mathcal{Z}_\infty}[Q_{K_{4+T}}] = \Omega(\log N)$*. We show the stronger statement that:*

$$\mathbb{E}_{\mathcal{Z}_N}[Q_{K_4}] \quad = \zeta^N(0, 1, 2, 3) \quad < 1/36$$

Since $\mathbb{E}_{\mathcal{Z}_\infty}[q_{K_4}] < 1$*, we have that* $\Pr_{\mathcal{Z}_\infty}[q_{K_4}] < 1$*.*[8] *However, the converse is not true. Although* $\mathbb{E}_{\mathcal{Z}_\infty}[q_{K_4+T}] = \Omega(\log N)$*, still* $\Pr[q_{K_4+T}] \leq \Pr_{\mathcal{Z}_\infty}[K_4]$*. Thus, if a* K_4 *occurs, then it is likely to have a large number of such tails.*

This example illustrates that to understand $\text{Th}(\mathcal{Z}_N)$, it is insufficient to consider expectation alone, and we must examine the structure of the query.

Before stating our general result, we describe how to tie together all of our tools to establish that oriented triangles are in $\text{Th}(\mathcal{Z}_\infty)$.

Example 3.4. *Let* q_{K_3} *be the following query:*

$$\begin{aligned} q_{K_3}() &= R(x, y), R(y, z), R(x, z), x < y, y < z \\ Q_{K_3}(x, y, z) &= R(x, y), R(y, z), R(x, z), x < y, y < z \end{aligned}$$

[8] It is easy to see that $\Pr[q_{K_4}] > 0$, which establishes that there is no zero-one law for $\mathcal{Z}_\infty$.

Query Shape	**Biggest Zeta**	**Asymptotic Expectation**
Path of length $t \geq 1$	$\zeta^N(0, \{1\}_t)$	$\Theta(N)$
Simple cycle of size $c \geq 3$	$\zeta^N(0, \{1\}_{c-1}, 2)$	$\Theta(\log N)$
Single cycle of length $c \geq 3$ with t other nodes	$\zeta^N(0, \{1\}_{c-1}, 2, \{1\}_t)$	$\Theta(\log^{t+1} N)$
Unification of the above	$\zeta^N(0, \{1\}_{c-1}, 2, \{1\}_{2t})$	$\Theta(\log^{2t+1} N)$
Unification of the above	$\zeta^N(0, \{1\}_{2c-2}, 2, 2, \{1\}_{2t})$	$\Theta(\log^{2t} N)$
Bicycle (b_1, b_2, t)	$\zeta(0, \{1\}_{b_1-1}, 2, \{1\}_{b_1+t-1}, 2)$	$\Theta(1)$

Figure 1: For each graph pattern, we give the largest Zeta function, by majorization, and the leading term of its expectations.

By Proposition 3.1, we have $\mathbb{E}_{\mathcal{Z}_\infty}[Q_{K_3}] = \zeta(0, 1, 2)$ and by Lemma 3.1, we have $\mathbb{E}_{\mathcal{Z}_\infty}[Q_{K_3}] = \Omega(\log N)$. To apply Janson's inequality, we need to describe the unifications of Q_{K_3} with itself that properly overlap. A quick check reveals that all such structures are images of the following query, $Q_1(x, y, z, u) =$

$$R(xy), R(yz), R(zx), R(yu), R(ux), x < y, y < z.$$

Using Proposition 3.4, a highest expectation ordering is $x < y < z < u$: let Q_1 with the additional comparison $z < u$ be Q_0. Then, using Proposition 3.1 and Lemma 3.1, $\mathbb{E}_{\mathcal{Z}_N}[Q_0] = \zeta^N(0, 1, 2, 2) = \mathcal{O}(1)$. As this is a highest expectation ordering and there is only a constant number of unifications (in N), this tells us that $\Delta = \mathcal{O}(1)$. Finally, Janson's inequality (Corollary 2.1) tells us that

$$\lim_{N \to \infty} \Pr_{\mathcal{Z}_N}[q_{K_3}] = 1, \textit{ i.e., } q_{K_3} \in \mathrm{Th}(\mathcal{Z}_\infty)$$

We describe $\mathrm{Th}(\mathcal{Z}_\infty, \mathrm{CQ}_{1B})$ in two steps. The first step is to decide membership in $\mathrm{Th}(\mathcal{Z}_\infty, \mathrm{CQ}_{1B}^{TO})$, in which the variables in the query are totally ordered. Then we will show that $q \in \mathrm{Th}(\mathcal{Z}_\infty, \mathrm{CQ}_{1B})$ if and only if we can find an ordering of the variables θ (i.e., a map to CQ^{TO}) such that $\theta(q) \in \mathrm{Th}(\mathcal{Z}_\infty, \mathrm{CQ}_{1B}^{TO})$.

Step 1: Deciding $q \in \mathrm{Th}(\mathcal{Z}_\infty, \mathrm{CQ}_{1B}^{TO})$. We first consider CQ_{1B}^{TO}, and we are able to give an explicit characterization of $\mathrm{Th}(\mathcal{Z}_\infty, \mathrm{CQ}_{1B}^{TO})$. Let $G(q) = (V, E)$ be the directed graph naturally associated to a query in which the variables are nodes and two nodes are connected if the variables to which they correspond participate in some subgoal in q.

Proposition 3.5. *For $q \in \mathrm{CQ}_{1B}^{TO}$, $q \in \mathrm{Th}(\mathcal{Z}_\infty)$ if and only if $G(q)$ if each connected component contains at most one cycle.*

We outline the main steps of the proof. In the forward direction, we use the expectations in Figure 1 and Janson's inequality (Corollary 2.1) as we have done in Example 3.4. We show the reverse direction by explicitly computing the probability of a family of queries, which we call bicycle graphs, that are essentially two cycles connected by a path. We show that every member of this family has probability less than 1. Then, the proposition follows by observing that an image of one member of the bicycle family can be embedded in any graph with two cycles. This implies that, for any query in $q \in \mathrm{CQ}_{1B}^{TO}$ with two cycles, $q \notin \mathrm{Th}(\mathcal{Z}_\infty)$.

Step 2: Deciding $q \in \mathrm{Th}(\mathcal{Z}_\infty, \mathrm{CQ}_{1B})$. We first describe the challenge to extend our result to CQ_{1B}. We have:

$$\Pr[q] = \Pr\left[\bigvee_{\theta \in \mathrm{Hom}(q, \mathrm{CQ}^{TO})} \theta(q)\right]$$

That is, we can reduce evaluation of q to the problem of evaluating the disjunction of several queries in CQ^{TO}. It's tempting to take a union bound, i.e., by writing

$$\Pr_{\mathcal{Z}_\infty}\left[\bigvee_{\theta \in \mathrm{Hom}(q, \mathrm{CQ}^{TO})} \theta(q)\right] \leq \sum_{\theta \in \mathrm{Hom}(q, \mathrm{CQ}^{TO})} \mathbb{E}_{\mathcal{Z}_\infty}[\theta(Q)].$$

For some queries, this bound is tight enough. But in general, the right-hand side may be much bigger than 1 and so the bound is trivial—even if $\Pr[\theta(q)] < 1$ for each θ. The observation is that $\theta(q)$ and $\theta'(q)$ are positively correlated as probabilistic events. In turn, we have:

$$\Pr_{\mathcal{Z}_\infty}\left[\bigvee_{\theta \in \mathrm{Hom}(q, \mathrm{CQ}^{TO})} \theta(q)\right] \leq 1 - \left(1 - \prod_{\theta \in \mathrm{Hom}(q, \mathrm{CQ}^{TO})} \theta(q)\right)$$

Since $|\mathrm{Hom}(q, \mathrm{CQ}^{TO})|$ is constant with respect to N, we can conclude that $\Pr_{\mathcal{Z}_\infty}[\bigvee_{\theta \in \mathrm{Hom}(q, \mathrm{CQ}^{TO})} \theta(q)] < 1$ if and only if $\Pr_{\mathcal{Z}_\infty}[\theta(q)] < 1$ for each $\theta \in \mathrm{Hom}(q, \mathrm{CQ}^{TO})$.

This gives an exponential time algorithm, by enumerating all $\theta \in \mathrm{Hom}(q, \mathrm{CQ})$. More precisely, this is an NP-algorithm, since we only need to guess a single θ and check whether the resulting $\theta(q) \in \mathrm{Th}(\mathcal{Z}_\infty, \mathrm{CQ}_{1B}^{TO})$; this check can be done efficiently using Prop 3.5. The following proposition shows that it is unlikely that one could design a substantially more efficient algorithm.

Proposition 3.6. *Given a query q, it is NP-Complete to decide if $q \in \mathrm{Th}(\mathcal{Z}_\infty, \mathrm{CQ}_{1B})$.*

The proof relies on the combinatorial information from Proposition 3.5, namely that q is in the Theory if and only if there is an automorphism to a cycle-like structure. Deciding this efficiently would give an efficient algorithm to solve 3-coloring. It is worth noting that some special cases can be solved efficiently. For example, if q does not contain any comparison predicates, then checking this condition is trivially true (since $h(x) = 1$ for

$x \in \mathbf{var}(q)$ is a valid homomorphism). Also, if there are no self joins, or if there are not-equal constraints between each pair of variables, then using a greedy algorithm based on majorization (Proposition 3.4) gives a linear time algorithm to decide if $q \in \mathrm{Th}(\mathcal{Z}_\infty)$.

Discussion. While our techniques enable us to compute $\Pr_{\mathcal{Z}_\infty}[q]$ for $q \in \mathrm{CQ}$, we were unable to extend Proposition 3.5 to describe the Theory, as our current technique requires an explicit description of the two classes.

4. APPLICATION: $\mathrm{Th}(\mathsf{CHKNS}_\infty, \mathrm{CQ}_{1B})$

We first compute some descriptive statistics of the CHKNS model, e.g., the marginal probability of an edge or set of edges. Then we use the classical idea of sandwiching [3], where we ensure that the probability a query is true on CHKNS is in between two different scalings of $\mathcal{Z}_\infty$.

Recall the definition of CHKNS from Section 1; there is a parameter $\delta \in [0, 1]$ that controls a coin flip to add an edge (or not) as each node is introduced. In this section, we consider $\delta = 1$ to simplify our discussion, and we return to the case when $\delta \in (0, 1)$ in the extensions. (Note that $\delta = 0$ implies that the graph is empty).

4.1 Propositional Queries

We develop the tools to write tractable expressions for the probability of propositional queries, e.g.,

$$\Pr_{\mathsf{CHKNS}_N}[R(3, 4), R(4, 9), R(10, 11)].$$

In many models, e.g., in Zeta graphs or Erdös–Rényi graphs, computing this probability is trivial; in contrast, it is non-trivial in the CHKNS model. It is, however, easy to write an explicit equation for the probability that an edge occurs, denoted $\Pr_{\mathcal{C}_N}[R(i, j)]$, but with $\approx N$ terms:

$$1 - \Pr_{\mathsf{CHKNS}_N}[R(i, j)] = \prod_{n=u}^{N} \left(\frac{n(n-1)}{n(n-1)+1} \right) \qquad (2)$$

where $u = \max\{i, j\}$. This is a directed analog of CHKNS's model described in the introduction. The rationale for this product is that at instant n, there are n nodes and hence n^2 possible edges. And in the previous $n-1$ steps, we have already picked $n-1$ to add into the model. Hence, there are $n^2 - (n-1) = n(n-1)+1$ edges remaining. We want to avoid a particular edge $R(i, j)$, hence we want to pick one of the $n(n-1)$ edges. Thus, we have the fraction for each term. We then take a product over all such terms to arrive at Equation 2.[9]

This product is difficult to work with, as it involves a number of terms that depends on N. We resolve this using a standard technique to resolve infinite products in terms of Euler's Γ function.[10] To that end, we define four functions, r_-, r_+, U, and Z, as follows:

$$\begin{aligned} r_\pm(m) &= \frac{1}{2}\left(1 \pm \sqrt{1+4m}\right) \\ U(l, u, m) &= \frac{\Gamma(u - r_+(m))\Gamma(u - r_-(m))}{\Gamma(l - r_+(m))\Gamma(l - r_-(m))} \\ Z(l, u) &= \frac{\Gamma(l - r_+(-1))\Gamma(l - r_-(-1))}{\Gamma(u - r_+(-1))\Gamma(u - r_-(-1))} \end{aligned}$$

The pair of functions $r_\pm(m)$ are the roots of the polynomial $n^2 - n - m$, which we use to factor the numerator and denominator of Equation 2. Specifically, with this notation, we can rewrite Equation 2:

$$1 - \Pr_{\mathcal{C}_N}[R(i, j)] = Z(u, N+1)U(u, N+1, 0) \qquad (3)$$

This statement can be verified from the functional equation that defines the Γ function, namely, $\Gamma(z+1) = z\Gamma(z)$ and the above observation about $r_\pm(m)$. We can give a more general interpretation:

> $Z(l, u+1)U(l, u+1, m)$ *is the probability that none of* $m+1$ *distinguished edges is selected in any of the rounds between* l *and* u.

This formula allows us to compute the probability that one of several edges occurs (that may arrive at different times). We illustrate the idea by example.

Example 4.1. *Given* $R(a_i, b_i)$ *where for* $i = 1, \ldots, k$, $a_i, b_i \in [N]$ *such that the pairs* (a_i, b_i) *are distinct, let* $u_i = \max\{a_i, b_i\}$. *Suppose, without loss of generality, that* $u_1 \leq u_2 \leq \cdots \leq u_k$. *Then,* $\Pr[\bigvee_{i=1}^k R(a_i, b_i)] =$

$$1 - Z(u_1, N+1) \left(\prod_{i=1}^{k-1} U(u_i, u_{i+1}, i-1) \right) U(u_k, N+1, k-1)$$

This simply computes that each edge must be missing in the rounds when it is possible to be selected, e.g., an edge i *can only be selected if both end points are present, which implies that* $N \geq u_i$.

Using inclusion–exclusion, we can extend the computations in this example to compute the probability of any propositional expression of edges of graphs.

4.2 Asymptotic Edge Probabilities and an Intermediate Model

An advantage of this formulation is that since the Γ function is well studied, so we have standard approximations for Γ [15, p. 482]. Using these approximations, the following proposition is straightforward:

Proposition 4.1. *Let* $u = \max\{i, j\}$ *and* N *large enough,*

$$\Pr_{\mathit{CHKNS}_N}[R(i, j)] \in [u^{-1}, u^{-1} + u^{-2}/2]$$

Figure 2 shows several values for CHKNS_∞ and u of this function with indicators for Zeta graphs.

Proposition 4.1 offers a glimpse of the connection between Zeta graphs and the CHKNS model. To make

[9] Extending to the undirected case is straightforward and is therefore omitted.

[10] Euler's gamma function is defined as $\Gamma(t) = \int_0^\infty x^{t-1}e^t \, \mathrm{d}x$. For integer t, $\Gamma(t+1) = t!$, where $t!$ is the factorial function.

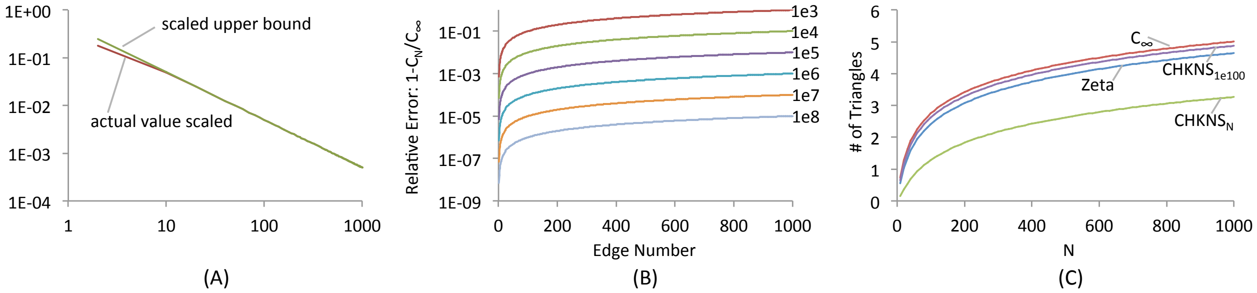

Figure 2: (A) The x axis is u and the red line plots $\log u \Pr_{\mathcal{C}_\infty}[R(i,j)]$ where $u = \max\{i,j\}$. Thus, 0 would be a perfect agreement with the lower bound of u from Proposition 4.1. We also plot the upper bound from Proposition4.1 in green. Visually, the upper bound is in close agreement with the true value. Together (B) and (C) give a glimpse at the two ways CHKNS differs from $\mathcal{C}_\infty$: (B) shows the effect of edge probability growing as N increases, while (C) illustrates the negative correlations in CHKNS (versus the edge-independent $\mathcal{C}_\infty$). (B) We calculate the relative error in using CHKNS$_N$ (versus C_∞) to compute the probability of an edge for the first 1000 edges as we vary N in CHKNS$_N$: for an edge (i,j) let $u = \max\{i,j\}$; for a fixed N we vary u along the x axis and plot $1 - \frac{\Pr_{\mathsf{CHKNS}_N}[R(i,j)]}{\Pr_{\mathcal{C}_\infty}[R(i,j)]}$ for choices of N. (C) Shows the expected number of triangles ($\mathbb{E}_X[Q_{K_3}]$) that occur within the first N nodes, where $X \in \{\mathcal{Z}_\infty, \mathcal{C}_\infty, \mathsf{CHKNS}_N\}$. The rate of growth is approximately the same for all models. However, we can see a small difference between $\mathcal{C}_\infty$ and CHKNS even at very large N values (10^{100}) due to correlation. This suggests that Proposition 4.3 is loose for larger values of N; an observation that we leverage in our final result.

the connection precise, we introduce a model $\mathcal{C}_\infty$ that is an edge-independent version of CHKNS. That is, each tuple $R(i,j)$ is an independent event with probability $\Pr_{\mathcal{C}_\infty}[R(i,j)]$ defined as

$$\Pr_{\mathcal{C}_\infty}[R(i,j)] = \lim_{N\to\infty} \Pr_{\mathsf{CHKNS}_N}[R(i,j)]$$

The probability of an edge in $\mathcal{C}_\infty$ is its limit in CHKNS.

Below, we use $\mathcal{C}_\infty$ to compute $\mathrm{Th}(\mathsf{CHKNS}_\infty, \mathrm{CQ}_{1B})$. First, we establish the relationship between $\mathrm{Th}(\mathcal{C}_\infty, \mathrm{CQ}_{1B})$ and $\mathrm{Th}(\mathcal{Z}_\infty, \mathrm{CQ}_{1B})$. To this end, we observe that the edge probabilities do not change too much between the two models:

Corollary 4.1. *For $i, j \geq 1$, we have*

$$\frac{\Pr_{\mathcal{C}_\infty}[R(i,j)]}{\Pr_{\mathcal{Z}_\infty}[R(i,j)]} \in [1, 1.5] \text{ and } \Pr_{\mathcal{C}_\infty}[R(i,j)] \leq \Pr_{\mathcal{Z}_\infty}[R(i{+}1, j{+}1)]$$

Since expectation is linear in these quantities, the forward direction of Proposition 3.5 is immediate (since $\Pr_{\mathcal{Z}_\infty}[q] \leq \Pr_{\mathcal{C}_\infty}[q]$ for any monotone q). In the reverse direction, we observe that for every connected query $q \notin \mathrm{Th}(\mathcal{Z}_\infty, \mathrm{CQ}_{1B}^{TO})$, we have $\mathbb{E}_{\mathcal{Z}_N}[Q] < 0.5$. Thus, the second portion of the above corollary implies that we can bound term-by-term by simply mapping each term in the expectation to $c_i' = c_i + 1$. The only terms not covered by this mapping contribute a vanishingly small amount. In particular, a loose bound is $\mathbb{E}_{\mathcal{C}_\infty}[Q] < 0.75$. The rest of the reasoning for Proposition 3.5 is unchanged. Therefore, we have

Proposition 4.2. *With the notation above,*

$$\mathrm{Th}(\mathcal{C}_\infty, \mathrm{CQ}_{1B}) = \mathrm{Th}(\mathcal{Z}_\infty, \mathrm{CQ}_{1B})$$

4.3 Upper and Lower Models

We show that $\mathrm{Th}(\mathcal{C}_\infty, \mathrm{CQ}_{1B}^{TO}) = \mathrm{Th}(\mathsf{CHKNS}_\infty, \mathrm{CQ}_{1B}^{TO})$. First, we observe that:

$$\Pr_{\mathsf{CHKNS}}[q] \leq \Pr_{\mathcal{C}_\infty}[q] \text{ for } q \in \mathrm{CQ}$$

Thus, $\mathrm{Th}(\mathcal{C}_\infty, \mathrm{CQ}_{1B}) \supseteq \mathrm{Th}(\mathsf{CHKNS}_\infty, \mathrm{CQ}_{1B})$. The proof is straightforward: Proposition 4.1 shows that every edge is upper bounded by its probability in $\mathcal{C}_\infty$ and that the edges in CHKNS are negatively correlated (formally, see Proposition 4.3).

Establishing that $\mathrm{Th}(\mathcal{C}_\infty, \mathrm{CQ}_{1B}) \subseteq \mathrm{Th}(\mathsf{CHKNS}_\infty, \mathrm{CQ}_{1B})$. Establishing the reverse inequality will take essentially the remainder of this section. The first issue is to handle the correlations between edges in CHKNS. We establish that any positive conjunction of k propositions is within a constant factor, γ on C_∞.

Proposition 4.3. *For $N \geq 0$, let $e_1, \ldots, e_k$ be distinct edges (i.e., elements of $[N]^2$), then*

$$\Pr_{\mathcal{C}_\infty}\left[\bigwedge_{i=1}^{k} e_i\right] \geq \Pr_{\mathit{CHKNS}_\infty}\left[\bigwedge_{i=1}^{k} e_i\right] \geq \gamma \prod_{i=1}^{k} \Pr_{\mathcal{C}_\infty}[e_i] \text{ and}$$

$$\gamma^{-1} \Pr_{\mathcal{C}_\infty}\left[\bigvee_{i=1}^{k} e_i\right] \geq \Pr_{\mathit{CHKNS}_\infty}\left[\bigvee_{i=1}^{k} e_i\right] \geq \prod_{i=1}^{k} \Pr_{\mathcal{C}_\infty}[e_i]$$

in which $\gamma = \frac{1}{\Gamma((3+\sqrt{3})/2)((3-\sqrt{3})/2)} \approx 0.581$ *and so* $\gamma^{-1} \approx 1.721$.

This proposition shows that one can lower bound the CHKNS model with $\mathcal{C}_\infty$. The proof is by direct calculation and approximates the inclusion–exclusion terms.

A consequence of the above result is that the expectation of any full query is within constant factors on $\mathcal{C}_\infty$ and $\mathcal{Z}_\infty$. Since all queries in CQ_{1B}^{TO} that are in $\mathrm{Th}(\mathcal{Z}_\infty)$ have unbounded expectation, this implies

$$\mathrm{Th}(\mathcal{C}_\infty, \mathrm{CQ}_{1B}^{TO}) \supseteq \mathrm{Th}(\mathsf{CHKNS}_\infty, \mathrm{CQ}_{1B}^{TO}).$$

In the other direction, we use Proposition 4.3 to see that all those not in the theory have expectation less than 0.75. Thus, we have shown:

Proposition 4.4. *With respect to* CQ_{1B}^{TO} *all three Theories are equal.*

$$\mathrm{Th}(\mathcal{Z}_\infty, \mathrm{CQ}_{1B}^{TO}) = \mathrm{Th}(\mathcal{C}_\infty, \mathrm{CQ}_{1B}^{TO}) = \mathrm{Th}(\mathit{CHKNS}_\infty, \mathrm{CQ}_{1B}^{TO})$$

We now observe that if $q \in \mathrm{Th}(\mathcal{C}_\infty, \mathrm{CQ}_{1B})$ then there is some mapping $\theta \in \mathrm{Hom}(q, \mathrm{CQ}^{TO})$ such that $\Pr[\theta(q)] = 1$. This same mapping applies to CHKNS just as well, and by the above we have:

$$\mathrm{Th}(\mathcal{C}_\infty, \mathrm{CQ}_{1B}) \supseteq \mathrm{Th}(\mathsf{CHKNS}_\infty, \mathrm{CQ}_{1B}).$$

The reverse direction is more technically involved. To mirror the case of $\mathcal{C}_\infty$ (or our analysis of $\mathrm{Th}(\mathcal{Z}_\infty, \mathrm{CQ}_{1B})$), we would like to argue that if $\Pr[\theta(q)] < 1$ for each $\theta \in \mathrm{Hom}(q, \mathrm{CQ}^{TO})$ then

$$\Pr_{\mathsf{CHKNS}_\infty}\left[\bigvee_{\theta \in \mathrm{Hom}(q, \mathrm{CQ}^{TO})} \theta(q)\right] < 1$$

But we only know $\Pr_{\mathsf{CHKNS}_\infty}[\bigvee_{\theta \in Hom(q,\mathrm{CQ}^{TO})} \theta(q)] < 1 - \varepsilon$, where ε is a constant $\varepsilon > 0$ that may depend on q but does not depend on N. From Proposition 4.3, this argument may give us a meaningless upper bound if $(1 - \varepsilon)/\gamma \geq 1$. The key technical observation is that γ is very loose. In particular, in the statement of Proposition 4.3, if $u_* = \min_i u_i$ where $u_i = \min\{e_{i_1}, e_{i_2}\}$ then one can show:

$$\left(1 + \mathcal{O}(u_*^{-1})\right) \Pr_{\mathsf{C}_\infty}\left[\bigvee_{i=1}^{k} e_i\right] \geq \Pr_{\mathsf{CHKNS}_\infty}\left[\bigvee_{i=1}^{k} e_i\right]$$

Motivated by this observation, we introduce a class of models $\mathcal{C}_{l,N}$ that has a node set $[l, N] = \{l, \ldots, N\}$ and for $i, j \in [l, N]$ we have

$$\Pr_{\mathcal{C}_{l,N}}[R(i,j)] = \Pr_{\mathcal{C}_N}[R(i,j)]$$

We then show that for any $l \geq 1$,

$$\mathrm{Th}(\mathcal{C}_{l,\infty}, \mathrm{CQ}_{1B}) = \mathrm{Th}(\mathcal{C}_\infty, \mathrm{CQ}_{1B})$$

To establish this statement, we make a detour back to Zeta graphs and prove that the analogous statement holds for Zeta graphs. That is,

$$\mathrm{Th}(\mathcal{Z}_{l,\infty}, \mathrm{CQ}_{1B}) = \mathrm{Th}(\mathcal{Z}_\infty, \mathrm{CQ}_{1B})$$

where $\mathcal{Z}_{l,\infty}$ is defined in the obvious way. The technical issue here is to observe that Lemma 3.1 holds even if we slice off all terms below l. For sums like $\sum_{i=l}^{N} i^{-1} = \Theta(\log N)$, our statement is trivial: removing only a finite number of terms from a divergent series keeps the series diverging at the same rate. However, for sums like

$$\zeta_a^N(1,1) = \sum_{a \leq i_1 < i_2 \leq N} i_1^{-1} i_2^{-1}$$

we are removing infinitely many terms from $\zeta^N(1,1)$. The proof of Lemma 3.1 needs this stronger result.

Using the family of graphs $C_{l,\infty}$, we are able to show our main result for this section. Essentially, given any q, we pick l so that the argument we outline works. Summarizing our discussion,

Proposition 4.5. *For each* $l \geq 0$

$$\mathrm{Th}(\mathcal{C}_{l,\infty}, \mathrm{CQ}_{1B}) = \mathrm{Th}(\mathit{CHKNS}, \mathrm{CQ}_{1B})$$

We now turn to extending this result when $\delta \in (0,1)$.

Extending to CHKNS ***with*** $\delta \in (0,1)$**.** We describe how to extend our results to the case in which δ is a constant in $(0,1)$. We denote this as $\mathsf{CHKNS}_{N,\delta}$ with the obvious meaning. Note that the case $\delta = 0$ is the empty graph, and its theory is straightforward. It is also straightforward to show that for $\delta < 1$ there are fewer edges than when $\delta = 1$, and so fewer queries in the Theory. To show the reverse direction requires some work: First, we observe that $\mathbb{E}_{\mathsf{CHKNS}_{\delta,\infty}}[Q] \approx \delta^k \mathbb{E}_{\mathsf{CHKNS}_{1,\infty}}[Q]$, where k is the number of relations in the body of Q (where $q \in \mathrm{CQ}^{TO}$). Since our arguments based on Janson's inequality arguments depend on whether $\mathbb{E}[Q]$ is bounded or not, this result is enough to prove that all queries in $\mathrm{Th}(\mathsf{CHKNS}, \mathrm{CQ}_{1B})$ are in $\mathrm{Th}(\mathsf{CHKNS}_{N,\delta})$ (we still have to go through the same technique we used for $\mathcal{C}_{l,\infty}$). Our techniques do not give information about the global structure (e.g., whether there is a single, giant component) as CHKNS's methods do; instead they tell us about the fine structure of these larger structures.

To establish that $\mathbb{E}_{\mathsf{CHKNS}_{\delta,\infty}}[Q] \approx \delta^k \mathbb{E}_{\mathsf{CHKNS}_{1,\infty}}[Q]$, we need to redo the technical tools in this section. One added technical complication is that for $\delta \in (0,1)$, the probability distribution over edges depends on how many edges have been introduced in previous steps; in contrast at $\delta = 1$, there are always t edges at time t. To cope, we show that a fractional version of Equation 2 is within a small constant (vanishing in u) of the above Markov chain. The remaining techniques apply in a straightforward but tedious way.

5. RELATED WORK

There is a staggering amount of work on network analysis. We refer readers to two textbooks in this area: Wasserman and Faust [23] and Newman's introduction to networks [21]. There are also several recent surveys as this area continues to explode in interest. We do not hope to completely summarize these areas, but rather to describe the work that is most technically relevant to this paper.

Arguably one of the most famous works on random network graph models is the Albert and Barabási preferential attachment model [2]. Neither CHKNS's model nor Zeta graphs are preferential attachment models; in particular, neither is scale free. In Albert and Barabási's seminal work [2, p. 73], this is the model they call Model A, which contains half of the preferential attachment story. It is interesting future work to incorporate the additional information of preferential attachment. One simple observation is that rather than selecting a match to each node, one could view this model as selecting links to attach to. Thus, the relaxation of this graph may be some kind of dual to Zeta graphs. We are also able to show that one preferential attachment graph (where the probability two edges connected is proportional to the degree of a node) is actually a slight generalization of zeta graph, and we are able to extend much of our theory to such graphs. However, it has been observed that several important properties of real network graphs are not captured by these models, and researchers are designing higher fidelity models. Still, there is no consensus on the right model and all known models capture some aspects, but fail to capture others. For example, there are currently no models that match the hyperbolicity or clustering of real-world graphs [7, 21]. Our work here explicitly does not argue about whether a particular model has higher fidelity with a particular empirical aspect of real-world graphs; instead our work is about using databse theory to contribute to the theoretical underpinning of these models.

Theoretical researchers have approximated these distributions to prove theorems. For example, most closely related is Lynch [19], who captured the power-law distribution by allowing a distribution over all nodes with a specified degree sequence. He showed that such graphs exhibited a zero-one law for first-order logic. However, as we show, there are conjunctive queries (without constants, of course) for which CHKNS's model does not have a zero-one law.

Our model is close to an Erdös–Rényi model, and so it's not surprising that our techniques borrow from query answering on Random graphs, e.g., Dalvi et al.'s work [10]. Our results build on their results, but our technical challenges are different: the bulk of work in this paper goes into dealing with correlations, inhomogenous probability values, and the technical difficulties that have to do with inifinite series to compute simple propositional expressions—none of these challenges are present in Dalvi et al.'s work. The complexity of related questions seems to be higher in the model here. For example, in Dalvi et al.'s work, one source of hardness is unification, e.g., if the automorphism group of q is trivial, then it is not hard to show that their algorithm is in P-time. However, in our setting, ♯P-hardness holds even for families of queries with a trivial automorphism group, e.g., if all relational symbols in q are distinct.

Shelah and Spencer [22] gave a nearly complete classification of sparse, Erdös–Rényi random graphs. If we examine Zeta graphs, we see that the expected number of edges in Z_N is $\sum_{i=1} \sum_{j \leq i} i^{-1} = N$. Hence, a natural ER modeling is that $p = N^{-1}$. However, in $G(N, N^{-1})$ model notice that the theory is slightly different: each node has a constant expected degree, while any fixed node in Zeta graphs has unbounded degree as $N \to \infty$. In some ways, our theory is closer to $G(N, p)$ $p = \frac{\log N}{N}$. These are further indications that the class of random graph models we consider do not coincide with these previous models.

6. CONCLUSION AND FUTURE WORK

We studied the theory of graph queries on two random graph models. We described Zeta databases, which were motivated by our desire to find a simple Erdös–Rényi-like model that would allow us to answer traditional database-theory questions. Our technical contributions were the basic tools for conjunctive query answering on Zeta databases, and a complete characterization of the Theory for a language inspired by conjunctive graph patterns (CQ_{1B}). These techniques were simplified by a well-developed set of tools for dealing with multiple-valued zeta functions, which have been developed for a completely unrelated purpose.

Our future work will be in two directions: (1) *more expressive languages* and (2) *higher-fidelity graph models*. For (1), to extend our results beyond CQ_{1B} to all of CQ, and perhaps all of first-order logic, to more fully compare our results with those of Lynch [19]. It may also be interesting to investigate the introduction of constants into the language. For (2), we plan to add in constraints on preferential attachment following Albert and Barabási's work [2]. One interesting result here would be an analytic theory of the statistical signficance of motifs for this popular family of random graphs.

Acknowledgments. The author would like to thank the anonymous ICDT reviewers for both their careful reviews and constructive comments. The author also would like to thank Eytan Adar for helpful discussions about graph models and pointing us toward Newman's book [21], which provided a wonderful introduction to the area. The author also thanks Ce Zhang and Victor Bittorf for help in preparing the figures and code for this project. CR's work on this project is generously supported by NSF CAREER Award under No. IIS-1353606 and a Sloan Research Fellowship. CR is additionally supported under NSF award No. CCF-1356918, ONR awards No. N000141210041 and No. N000141310129, the AFOSR, and gifts from Oracle, Google, and Lightspeed Ventures.

7. REFERENCES

[1] William Aiello, Fan R. K. Chung, and Linyuan Lu. Random evolution in massive graphs. In *FOCS*, pages 510–519, 2001.

[2] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, Jan 2002.

[3] Noga Alon and Joel Spencer. *The Probabilistic Method*. John Wiley, 1992.

[4] Johannes Blümlein and Stefan Kurth. Harmonic sums and mellin transforms up to two-loop order. *Phys. Rev. D*, 60:014018, Jun 1999.

[5] Graham Brightwell and Peter Winkler. Counting linear extensions is #p-complete. STOC '91, pages 175–181, New York, NY, USA, 1991. ACM.

[6] D. S. Callaway, J. E. Hopcroft, J. M. Kleinberg, M. E. J. Newman, and S. H. Strogatz. Are randomly grown graphs really random? *Phys Rev E*, 64(4):041902, October 2001.

[7] Wei Chen, Wenjie Fang, Guangda Hu, and Michael W. Mahoney. On the hyperbolicity of small-world networks and tree-like graphs. *CoRR*, abs/1201.1717, 2012.

[8] C. Costermans, J. Y. Enjalbert, Hoang Ngoc Minh, and M. Petitot. Structure and asymptotic expansion of multiple harmonic sums. In *ISSAC '05*, 2005.

[9] Richard E. Crandall. Fast evaluation of multiple zeta sums. *Math. Comput.*, 67(223), July 1998.

[10] Nilesh N. Dalvi, Gerome Miklau, and Dan Suciu. Asymptotic conditional probabilities for conjunctive queries. In *ICDT*, 2005.

[11] Herbert B. Enderton. *A mathematical introduction to logic*. Academic Press, 1972.

[12] Ronald Fagin. Probabilities on finite models. *J. Symb. Log.*, 41(1):50–58, 1976.

[13] Ronald Fagin, Benny Kimelfeld, and Phokion G. Kolaitis. Probabilistic data exchange. *J. ACM*, 58(4):15, 2011.

[14] Yu.V. Glebskii, D.I. Kogan, M.I. Liogon'kii, and V.A. Talanov. Range and degree of realizability of formulas in the restricted predicate calculus. *Cybernetics*, 5(2):142–154, 1969.

[15] Ronald L. Graham, Donald E. Knuth, and Oren Patashnik. *Concrete Mathematics: A Foundation for Computer Science*. Boston, MA, USA, 2nd edition, 1994.

[16] Michael Hoffman. *References on Multiple Zeta Values and Euler Sums*. `http://www.usna.edu/Users/math/meh/biblio.html`, November 2012.

[17] Jure Leskovec, Deepayan Chakrabarti, Jon M. Kleinberg, Christos Faloutsos, and Zoubin Ghahramani. Kronecker graphs: An approach to modeling networks. *Journal of Machine Learning Research*, 11:985–1042, 2010.

[18] Leonid Libkin. *Elements of Finite Model Theory*. Springer, 2004.

[19] James F. Lynch. Convergence law for random graphs with specified degree sequence. *ACM Trans. Comput. Log.*, 6(4):727–748, 2005.

[20] M. E. J. Newman, D. J. Watts, and S. H. Strogatz. Random graph models of social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(Suppl 1):2566–2572, 2002.

[21] Mark Newman. *Networks: An Introduction*. New York, NY, USA, 2010.

[22] Saharon Shelah and Joel Spencer. Zero-one laws for sparse random graphs. *Journal of the American Mathematical Society*, 1(1), 1988.

[23] S. Wasserman and K. Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[24] V. V. Zudilin and W. Zudilin. Algebraic relations for multiple zeta values. *Russian Mathematical Surveys*, 58:1–29, 2003.