ON THE EXPECTED SIZE OF RECURSIVE DATALOG QUERIES

by

S. Seshadri and Jeffrey F. Naughton

Computer Sciences Technical Report #1019

March 1991

On the Expected Size of Recursive Datalog Queries *

S. Seshadri and Jeffrey F. Naughton
{seshad, naughton}@cs.wisc.edu
Department of Computer Sciences
University of Wisconsin-Madison

Abstract

We present asymptotically exact expressions for the expected sizes of relations defined by three well-studied Datalog recursions, namely the "transitive closure", "same generation" and "canonical factorable recursion". We consider the size of the fixpoints of the recursively defined relations in the above programs, as well as the size of the fixpoints of the relations defined by the rewritten programs generated by the Magic Sets and Factoring rewriting algorithms in response to selection queries. Our results show that even over relatively sparse base relations, the recursively defined relations are within a small constant factor of their worst-case size bounds, and that the Magic Sets rewriting algorithm on the average produces relations within a small constant factor of the corresponding bounds for the recursion without rewriting. The expected size of relations produced by the Factoring algorithm, when it applies, is significantly smaller than the expected size of relations produced by Magic Sets. This lends credence to the belief that reducing the arity of the recursive predicate is probably more important than restricting the recursion to relevant tuples.

1 Introduction

A great deal of work has been done on recursive queries and query rewriting techniques. Surprisingly, however, almost nothing is known about the expected-case behavior of these queries or their rewritten versions. In this paper, we derive analytic expressions for the sizes of the fixpoints of the recursively defined relations in the "transitive closure", "same generation" and the "canonical factorable recursion" Datalog programs. Similarly, we derive analytic expressions for the size of the fixpoints of relations defined by the rewritten programs generated by the Magic Sets and Factoring rewriting algorithms in response to selection queries over these Datalog programs. Our experimental results show close agreement with the analytic formulas even for small base relation sizes.

*This work was supported by NSF grant IRI-8909795 and by a grant of the Wisconsin Alumni Research Foundation.

The actual programs for the three recursions that we consider are presented in Section 2. The transitive closure is a basic and well-studied Datalog recursion. We have chosen the same generation query because it is simple, yet it gives a lot of insight into more complex recursions. We have chosen the canonical factorable recursion because it is representative of an important class of easily evaluable recursions, the factorable recursions.

The Magic Sets rewriting technique attempts to use a selection query to restrict the fixpoint evaluation of the recursion to search and generate only those portions of the relations that are relevant to the query. The Factoring rewriting technique, in addition to restricting the fixpoint evaluation, also reduces the arity (number of arguments) of the recursive predicate. The analytic expressions we derive give insight into how successful these techniques are over random base relations.

To investigate the expected behavior of these recursions, we view the (binary) base relations as random digraphs. That is, there is a tuple (x, y) in a relation if and only if there is a directed edge from x to y in the corresponding digraph. In the model of random digraphs we have adopted, denoted $D_{n,p}$, each edge of an n node digraph is present with probability p, independently of the presence or absence of other edges. The size of the recursive relations depends critically on np, the expected out-degree of a vertex. In the following, let a and b be the base relations for the programs we consider in Section 2. Let np be a constant $c > 1$, and let the digraphs corresponding to a and b be chosen randomly from $D_{n,c/n}$. (For clarity of exposition, we will assume that the out-degrees of the a and b digraphs are the same. The extension to different out-degrees is straightforward.) Also, let Θ be the unique root in $[0, 1]$ of the equation $1 - x - e^{-cx} = 0$, and let Θ' be the unique root in $[0, 1]$ of the equation $1 - x - (1 - \Theta)e^{-cx} = 0$. A summary of the key results appears in Table 1. TC refers to the transitive closure, SG refers to the same generation program and CFR refers to the canonical factoring program in the following table. More detailed results appear in later sections and a graph of these formulas vs. the constant c appears in Figures 3 and 4.

recursion	expected size (+ lower order terms)
TC (no rewriting)	$\Theta^2 n^2$
TC (magic sets)	$\Theta^4 n^2$
TC (factoring)	$\Theta^2 n$
SG (no rewriting)	$\Theta^2 n^2$
SG (magic sets)	$\Theta^4 n^2$
CFR (no rewriting)	$\Theta(2\Theta' - \Theta)n^2$
CFR (magic sets)	$\Theta^3(2\Theta' - \Theta)n^2$
CFR (factoring)	$\Theta(2\Theta' - \Theta)n$

Table 1: Summary of main results.

As noted above, the constant Θ ranges between zero and one. For $c = 2$ (recall that c is the average out-degree of the digraph), we have $\Theta = 0.799$, so $\Theta^2 = 0.638$ and $\Theta^4 = 0.408$. If we apply the Magic Sets transformation to evaluate a selection query of the form $q(Y)$:– $sg(1, Y)$ on digraphs with $c = 2$, Table 1 tells us that the relation materialized will on the

average contain $0.408n^2$ tuples. On 10,000 node digraphs (20,000 expected tuples in each of a and b), the relation materialized after rewriting by Magic Sets will thus on the average contain more than 40 million tuples.

While a statement about random relations does not imply anything about any specific "real world" instance of the problem, this result is worth noticing because it is in such stark contrast to the situation for non-recursive queries. For example, if we consider the join query $q2(Y)$:- $a(1,W), b(W,Y)$ on the same two digraphs , the expected answer size is four tuples. Furthermore, a selection-pushing evaluation strategy (in the presence of appropriate indices) will compute this answer with work proportional to the final answer size. As can be seen from Table 1, though the Magic Set Strategy buys you a factor of Θ^2 over the corresponding program without rewriting, as Θ grows (which corresponds to higher average out-degree for the underlying graphs), the gap between the two narrows until they are almost identical. For $c = 4$, we have $\Theta = 0.98$, $\Theta^2 = 0.96$ and $\Theta^4 = 0.923$ which means that the program produced by the Magic Sets rewriting strategy will materialize only 4% fewer tuples than the original program.

On the transitive closure and the canonical factorable recursion, Factoring does roughly a factor of n better than Magic Sets (Factoring does not apply to the same generation query). This gain of a factor of n is not due to better "focusing" properties of the rewritten factored program; rather, it is due to the fact that the Factoring strategy reduces the arity of the recursively defined predicate. Our results lend credence to the hypothesis, stated in [BKBR87], that using a selection to reduce the arity of the recursion is far more important than using the selection to avoid irrelevant tuples.

Related work on the performance of recursive queries and their evaluation algorithms [BMSU86, BR88, GKS91, HN88, HL86, MSPS87, Nau88, SZ87] has considered either worst-case performance, or performance over structured synthetic databases, or empirically measured performance over randomly generated relations. There is a vast and growing body of literature pertaining to random graph theory [Bol85]; the most closely related work to ours is a recent paper by Karp on the transitive closure [Kar90]. Many of the background results in Section 2 are taken from this paper.

The remainder of this paper is organized as follows. Section 2 develops necessary terminology and presents relevant previous results in random graph theory. Deriving our bounds on the expected sizes of relations required some new results about properties of random digraphs; we present these new results in Section 3. Section 4 gives our results about sizes of answers to the three recursions we consider, while Section 5 gives the results of empirical tests of our formulas. We conclude in Section 6.

2 Background

2.1 Recursions and Graph Problems

The three recursions we consider in this paper are the transitive closure, same generation, and the canonical factorable recursion.
1) In Datalog, the transitive closure can be written

```
tc(X,Y) :-  a(X,Z), tc(Z,Y).
tc(X,Y) :- a(X,Y).
```

We will call the above program TC henceforth. In response to a query $tc(1,Y)$?, the Magic Sets rewriting strategy [BMSU86, BR87, Ram88] will generate the program (which we call TC_{mg} henceforth)

```
m(1).
m(W) :- m(X), a(X,W).

mtc(X,Y) :- m(X), a(X,Z), mtc(Z,Y).
mtc(X,Y) :- m(X),a(X,Y).
```

The Factoring rewriting strategy [NRSU89] will produce the program (which we call TC_{factor} henceforth)

```
m(1).
m(W) :- m(X), a(X,W).

ftc(W) :- ftc(X), a(X,W).
```

Treating the base relation as a digraph, we can easily see that (x, y) is in the relation tc if and only if there is a path from x to y in the a graph. In the case of relation mtc, the additional constraint is that x should belong to the magic set (the relation m). Notice that the magic set is the reachability from 1 in the a graph. Finally, y belongs to the relation ftc if and only if there is a path from 1 to y in the a graph.
2) In Datalog, the same generation is written

```
sg(X,Y) :- a(X,W), sg(W,Z), b(Y,Z).
sg(X,X).
```

We will call the above program SG henceforth. In response to a query $sg(1,Y)$?, the Magic Sets rewriting strategy will generate the program (which we call SG_{mg} henceforth)

```
m(1).
m(W) :- m(X), a(X,W).

msg(X,Y) :- m(X), a(X,W), msg(W,Z), b(Y,Z).
msg(X,X) :- m(X).
```

The Factoring rewriting strategy does not apply to queries on the same generation. Treating the base relations as digraphs, we can easily see that (x, y) is in the relation sg if and only if there is a path from x to some vertex z_0 of length i in the a graph, and there is a path of the same length i from y to z_0 in the b graph. In the case of relation msg, the additional constraint is that x should belong to the magic set (the relation m). Notice that the magic set is the reachability from 1 in the a graph. Therefore if x is in the magic set, any vertex reachable from x is also in the magic set.
3) The canonical factorable recursion we consider is

```
t(X,Y) :- a(X,W), t(W,Y).
t(X,Y) :- t(X,Z), b(Y,Z).
t(X,X).
```

We will call the above program CFR (for canonical factorable recursion) henceforth. In response to the query $t(1, Y)$?, the Magic Sets rewriting strategy will produce the program (which we call CFR_{mg} henceforth)

```
m(1).
m(W) :- m(X), a(X,W).

mt(X,Y) :- m(X), a(X,W), mt(W,Y).
mt(X,Y) :- m(X), mt(X,Z), b(Y,Z).
mt(X,X) :- m(X).
```

The Factoring rewriting strategy will produce the program (which we call CFR_{factor} henceforth)

```
m(1).
m(W) :- m(X), a(X,W).

ft(Y) :- m(Y).
ft(Y) :- ft(Z), b(Y,Z).
```

For this query, (x, y) is in the relation t if and only if there is a path from x to some vertex z_0 in the a graph and there is a path from y to z_0 in the b graph. In the case of relation mt, the additional constraint is that x should belong to the magic set (the set of vertices reachable from 1 in the a graph). Finally, y belongs to ft if and only if there is a path from y to z_0 in the b graph, where z_0 belongs to the magic set.

We now have a graph theoretic formulation for the problem of finding the sizes of fixpoints of the relations. For example, to compute the size of the fixpoint of the relation t, we have to compute the number of vertex pairs (x, y) such that there is a path from x to some vertex z_0 in the a graph, and a path from y to z_0 in the b graph. Once we have this correspondence established, we can work with graphs rather than relations. Henceforth, we talk about graphs and the corresponding problems on graphs only.

2.2 Relevant Previous Results in Random Graph Theory

Consider a random digraph drawn from $D_{n,p}$. The following theorem demonstrates an important gap phenomenon: when $np = c$, where c is a constant greater than 1, the number of vertices reachable from a given vertex is very likely to be either very small (in the interval $[0, B \ln n]$) or very large (in the interval $[\Theta n - w(n)\sqrt{n}, \Theta n + w(n)\sqrt{n}]$).

Definition 2.1 $X(r)$ is defined to be the set of vertices reachable from vertex r in a directed graph. $Y(r)$ is defined to be the set of vertices that can reach vertex r in a directed graph. In other words $X(r)$ is the forward reachability and $Y(r)$ is the reverse reachability. A vertex is reachable from and can reach itself by definition. □

By symmetry, all the arguments in this paper hold for forward as well as reverse reachability.

Theorem 2.1 *[Kar90] Let c be a constant greater than 1. Let d be a positive constant. Let B be a constant greater than $(d+1)c(c-1)^{-2}$. Let $w(n)$ be a nondecreasing unbounded function. Let Θ be the unique root in [0,1] of the equation $1 - x - e^{-cx} = 0$. Let D be drawn from $D_{n,c/n}$. Then, $Pr[|X(r)| \notin [0, B \ln n] \cup [\Theta n - w(n)\sqrt{n}, \Theta n + w(n)\sqrt{n}]] < n^{-d}$ for all sufficiently large n.*

For digraphs drawn from $D_{n,c/n}$, where $c > 1$, the set of vertices reachable from vertex r is called *large* if $|X(r)|$ lies in the interval $[\Theta n - w(n)\sqrt{n}, \Theta n + w(n)\sqrt{n}]$, and *small* otherwise. The following theorem tells us that the probability that $X(r)$ is small tends to a constant that depends on c alone. Furthermore, the expected size of $X(r)$, given that $X(r)$ is small tends to another constant that also depends only on c.

Theorem 2.2 *[Kar90] Let D be drawn from $D_{n,c/n}$, where $c > 1$. Then as n tends to infinity,*

1. *the probability that $X(r)$ is small tends to $1 - \Theta$, where Θ is the unique root in $[0,1]$ of the equation $1 - x - e^{-cx} = 0$;*
2. *the expected size of $|X(r)|$, given that $X(r)$ is small , tends to $\frac{1}{1-c(1-\Theta)}$.*

The following theorem throws a lot of insight into the structure of a random digraph. It states that the fraction of vertices that have a large reachability in a random digraph is a constant that depends on c alone, plus an error term that goes to 0 as n goes to infinity. A similar result is shown for the fraction of vertices that have a large forward and reverse reachability.

Theorem 2.3 *[Kar90] Let* LARGEOUT *be the set $\{u|X(u)$ is large$\}$,* LARGEIN *be the set $\{v|Y(v)$ is large$\}$, and* LARGE *be the set $\{u|X(u)$ is large and $Y(u)$ is large$\}$. Let $w(n)$ be a non-decreasing unbounded function. Then, with probability tending to 1, $|LARGEOUT - \Theta n| < w(n)\sqrt{n \log n}$, and $|LARGEIN - \Theta n| < w(n)\sqrt{n \log n}$, and $|LARGE - \Theta^2| < w(n)\sqrt{n \log n}$.*

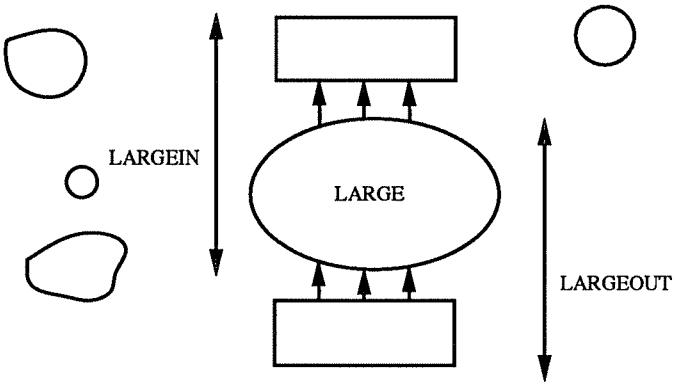

Figure 1: A Typical Random Digraph

Notice that the set *LARGE* is contained in the set *LARGEIN* as well as the set *LARGEOUT*.

The following is the fundamental Giant Strong Component Theorem for digraphs. It states that there will be exactly one strong component with more than $A \ln n$ vertices for some constant A. In fact, the proof shows that the set *LARGE* of the previous theorem corresponds to the Giant Strong Component.

Theorem 2.4 *[Kar90] Let $w(n)$ be a nondecreasing unbounded function. Let c be a constant greater than 1. There is a constant A such that with probability tending to 1, a digraph drawn from $D_{n,c/n}$ has exactly one strong component with more than $A \ln n$ vertices, and the number of vertices in that strong component differs from $\Theta^2 n$ by at most $w(n)\sqrt{n \log n}$.*

Roughly speaking, in all likelihood a graph drawn randomly from $D_{n,c/n}$ will have a huge strong component (*LARGE* of previous theorem) and its size will be $\Theta^2 n$. Apart from this the graph will have roughly $(\Theta - \Theta^2)n$ vertices that can reach some vertex in the strong component but are themselves not in the strong component. These vertices along with *LARGE* are the vertices of the set *LARGEOUT*. Similarly, there will be $(\Theta - \Theta^2)n$ vertices that can be reached from a vertex in *LARGE* but are themselves not in *LARGE*. These vertices along with *LARGE* are the vertices of the set *LARGEIN*. The rest of the $(2\Theta - \Theta^2)n$ vertices can neither reach a vertex in *LARGE* nor can be reached from *LARGE*. They are the ones that have a small forward and reverse reachability. Figure 1 shows how a typical random digraph looks.

The following theorem states that nearly all the vertices of a digraph drawn from $D_{n,c/n}$ can be expected to lie either in the largest strong component or in a strong component of size 1. In particular, almost certainly, all the cycles in the graph will lie in *LARGE*.

Theorem 2.5 *[Kar90] Let c be a constant greater than 1. Let Θ be the unique root in $[0,1]$ of the equation $1 - x - e^{-cx} = 0$. In digraph D, call a vertex exceptional if it lies in a strong component of size greater than 1 but is not in the unique largest strong component. Let the random variable s be the number of exceptional vertices in a digraph drawn from $D_{n,c/n}$. Then the expected value of s is bounded above by a function of n that converges to the positive constant $\frac{2c^2(1-\Theta)^2}{1-c(1-\Theta)} - \frac{c^2(1-\Theta)^4}{1-c(1-\Theta)^2}$.*

The following theorem shows that a random digraph contains a long simple cycle almost certainly. Note that the presence of a large strong component does not imply the existence of a long simple cycle.

Theorem 2.6 *[AKS81] Let c be a constant greater than 1. Let D be a digraph drawn from $D_{n,c/n}$. Then with probability tending to 1, there exists a simple cycle of length at least ϵn, $0 < \epsilon \leq 1$, in D.*

3 New Results On Random Digraphs

In this section we state our new results on random digraphs, which will play a key role in the next section when we derive analytic bounds on the sizes of fixponts of recursive relations. As before, let D be a typical graph drawn from $D_{n,c/n}$, $c > 1$, let $w(n)$ be an arbitrary nondecreasing unbounded function and let Θ be the unique root in $[0,1]$ of the equation $1 - x - e^{-cx} = 0$.

In the following theorem, we show that almost certainly there will be two cycles in the Giant Strong Component whose lengths are relatively prime.

Theorem 3.1 *For sufficiently large n, with probability tending to 1, there exist two cycles in the unique largest strong component of D of lengths l and k respectively such that $gcd(l,k) = 1$.*

Proof: Let $\phi(n)$ be the cardinality of the set $\{m \mid m < n \text{ and } gcd(m,n) = 1\}$. It was proved in [RS62] that a lower bound for $\phi(n)$ is $n/\log\log n$. This implies that given a cycle of length k, there are $k\phi(k)$ possible edges between non-consecutive vertices on the cycle whose presence would result in a cycle whose length is relatively prime to k. We will call these edges *distinguished edges.*

The idea behind the proof is to show that at least one of the distinguished edges for the long simple cycle mentioned in Section 2.2 will exist with overwhelming probability. For this purpose, we will view the graph D as being built in two stages. At the first stage will put in enough edges so that a big simple cycle is created in the graph with overwhelming probability. In the next stage we will add additional edges to create at least one of the distinguished edges for this simple cycle, again with overwhelming probability.

We will label the edges we add in the first stage "blue" and the edges we add at the second stage "red". In the final graph an edge from x to y exists if and only if there is either a blue edge or a red edge from x to y. The blue edges will be added with probability p_1 and the red edges with probability p_2. Therefore, the probability that there will neither be a blue edge nor a red edge between any two vertices is $(1-p_1)(1-p_2)$ which is $1-(p_1+p_2-p_1p_2)$. Therefore, if we choose $p_1 + p_2 = c/n$, the graph we obtain by the two stage process will have an edge probability smaller than c/n. So if the theorem holds for the graph obtained by the two stage process, then it will hold for a digraph drawn from $D_{n,c/n}$. Choose p_1 to be c_1/n for some $c_1 > 1$. Then, $p_2 = \frac{c-c_1}{n}$. After adding the blue edges alone the graph will have a cycle of

length ϵn by Theorem 2.6. By Theorem 2.5, with probability tending to 1, this cycle will lie in the unique largest strong component. Now add the red edges. The probability that none of the red edges added is one of the distinguished edges for the above cycle is $(1 - \frac{c-c_1}{n})^{\epsilon n \phi(\epsilon n)}$, which goes to zero as n tends to infinity. □

The following theorem derives an upper bound on the number of vertices that have short paths to a particular vertex.

Theorem 3.2 *Let $Z(r)$ be the number of vertices that have a path to r of length $\ln \ln n$ or less in D. Then, $Pr[Z(r) > n^{1/m}]$ tends to zero as n tends to infinity, for any constant $m > 0$.*

Proof: Let $W(i)$ stand for the in-degree of vertex i. Let α stand for $\ln \ln(n)$. Let k be the maximum of the in-degrees of any vertex in $Z(r)$. Then it is easy to see that $Z(r) \leq \frac{k^{\alpha+1}-1}{k-1}$. The second term in the previous inequality is the number of vertices in a tree where each vertex has in-degree k and there are $\alpha + 1$ levels. If $k \leq 1$, then $Z(r)$ is at most α and hence theorem holds. Assume $k > 1$ for the rest of the proof. $Z(r) > n^{1/m}$ implies $k^{\alpha+1} > n^{1/m}$. Therefore, $Z(r) > n^{1/m}$ implies $k > n^{\frac{1}{m(\alpha+1)}}$ and hence

$$\begin{aligned} & Pr[Z(r) > n^{1/m}] \\ \leq \quad & Pr[\exists x \mid x \in Z(r) \wedge W(x) > n^{1/m(\alpha+1)}] \\ \leq \quad & Pr[\exists x \mid W(x) > n^{\frac{1}{m(\alpha+1)}}] \\ \leq \quad & nPr[W(s) > n^{\frac{1}{m(\alpha+1)}}] \end{aligned}$$

We will now prove that $nPr[W(s) > n^{\frac{1}{m(\alpha+1)}}]$ goes to zero as n goes to infinity and the theorem will follow.

We will require the following bound on the tail of the binomial distribution [Rag86]. Let the random variable X have the distribution $BIN(n,p)$. Then, for every positive real β,

$$Pr[X > \beta np] < \left(\frac{e^{\beta-1}}{\beta^\beta}\right)^{np}$$

The in-degree of any vertex has the distribution $BIN(n, c/n)$. Hence the above inequality gives that $nPr[W(s) > n^{\frac{1}{m(\alpha+1)}}] < n\left(\frac{e^{\beta-1}}{\beta^\beta}\right)^c < (1/e)^c \frac{n}{(\beta/e)^{c\beta}}$, where $\beta c = n^{\frac{1}{m(\alpha+1)}}$. It can be verified that β grows faster than $\log n$ and that $(\log n)^{\log n}$ grows faster than n which implies that the above expression goes to zero as n tends to infinity. □

Theorem 3.3 *Let $Z(r)$ be the number of vertices that have a path from r of length $\ln \ln n$ or less in D. Then, $Pr[Z(r) > n^{1/m}]$ tends to zero as n tends to infinity, for any constant $m > 0$.*

Proof: The proof is similar to the previous theorem except that we will need to look at the distrbution of the out-degree of the vertices rather than their in-degree. □

We will now prove that given $\ln n$ random vertices, the probability that all of them have small reachability goes to zero as n tends to infinity.

Lemma 3.1 *Let U represent the event that X(u) is large and V represent the event that* $X(v_1)$, $X(v_2)$,...,$X(v_k)$ *are all small, where* $u, v_1, v_2 \ldots v_k$ *are distinct vertices in a graph D drawn from* $D_{n,c/n}$. *If* $k \leq \ln n$, *then* $Pr[U|V] \geq \Theta - O(\ln^3 n/n)$.

Proof: In [Kar90], a lemma similar to this one was proved. The difference is, the event V there denoted that the reachability of a single vertex was small. We adapt that proof to suit our lemma. Consider an experiment to determine, by a fanning out process, whether $X(u)$ is large. This experiment terminates as soon as $A \ln n$ vertices are reached. The experiment is unaffected by the information that $X(v_1)$,$X(v_2)$,... $X(v_k)$ are all small, unless a vertex in $X = X(v_1) \cup \ldots \cup X(v_k)$ is reached during the process. The probability that a random set $X(u)$ of size at most $A \ln n$ intersects X is at most $\frac{A \ln n|X|}{n}$. So, the probability that they do not intersect is at least $1 - \frac{A \ln n|X|}{n}$. Hence, we conclude that $Pr[U|V] \geq \Theta(1 - O(\ln^3/n))$, since $|X|$ is at most $O(\ln^2 n)$.

□

Corollary 3.1 *Let U represent the event that X(u) is small and V represent the event that* $X(v_1)$, $X(v_2)$,... $X(v_k)$ *are all small, where* $u, v_1, v_2 \ldots v_k$ *are distinct vertices in a graph D drawn from* $D_{n,c/n}$. *If* $k \leq \ln n$, *then* $Pr[U|V] \leq (1 - \Theta) + O(\ln^3 n/n)$.

Theorem 3.4 *Let* $A(u)$ *represent the event that* $X(u)$ *is small in D. Let* $v_1, v_2, \ldots, v_k$ *be k distinct vertices in D, where* $k = \ln n$. *Then* $Pr[A(v_1) \wedge A(v_2) \ldots A(v_k)]$ *goes to zero as n tends to infinity.*

Proof: Let $B(i)$ denote $A(v_i) \wedge A(v_{i+1}) \ldots \wedge A(v_k)$. Using Corollary 3.1, we get

$$
\begin{aligned}
& Pr[A(v_1) \wedge \ldots \wedge A(v_k)] \\
&= Pr[A(v_1)|B(2)]Pr[B(2)] \\
&\leq \{(1-\Theta) + O(\ln^3 n/n)\}Pr[A(v_2)|B(3)]Pr[B(3)] \\
&\vdots \\
&\leq \{(1-\Theta) + O(\ln^3 n/n)\}^k
\end{aligned}
$$

which goes to zero as n tends to infinity. □

The rest of this section will derive a theorem that is needed for the CFR programs alone. Let k_1 and k_2 be two positive constants such that $k_1 + k_2 = 1$, and let us randomly label $k_1 n$ of the vertices of a random digraph *type A* vertices and $k_2 n$ of the vertices *type B* vertices.

Definition 3.1 Let δ_r be a $0-1$ random variable defined for each vertex r as follows: $\delta_r = 1$, if $X(r)$ is small and $X(r)$ contains no *type A* vertex; $\delta_r = 0$, otherwise. □

Lemma 3.2 $Pr[\delta_r = 1]$ *tends to* $1 - \Theta'$ *where* Θ' *is the unique root in* $[0,1]$ *of the equation* $1 - x = k_2 e^{-cx}$.

Proof: If r is of *type A*, then $\delta_r = 0$. Therefore,

$$\begin{aligned} & Pr[\delta_r = 1] \\ = \; & Pr[\delta_r = 1 | r \text{ is not of } type\ A] Pr[r \text{ is not of } type\ A] \\ = \; & k_2 Pr[\delta_r = 1 | r \text{ is not of } type\ A] \end{aligned}$$

We will show that $Pr[\delta_r = 1 | r \text{ is not of } type\ A]$ is $\frac{1-\Theta'}{k_2}$ by a branching process argument.

The following argument is similar to one presented in [Kar90] for a simpler problem. Consider a fanning out process for constructing the set $X(r)$. The process constructs a sequence $\langle B_0, B_1 \ldots, B_t \ldots \rangle$, where the set B_i consists of vertices that have been reached during the first i iterations of the process. Here, $B_0 = \{r\}$ and $B_{i+1} = B_i \cup succ(v)$ where v is some vertex of B_i that has not been scanned so far (v is the vertex scanned at iteration $i+1$) and w lies in the set $succ(v)$ if and only if the digraph D contains the edge (v, w). The process terminates when, for some t, $B_t = t$; i.e., termination occurs when every vertex that has been reached has also been scanned. The number of vertices reached for the first time at iteration $i+1$ has the probability distribution $BIN(n-1-|B_i|, c/n)$. We will concentrate on the early stages of this process when $|B_i| \leq \ln n$, and then this distribution is closely approximated, for large n, by the probability distribution $BIN(n-1, c/n)$.

The evolution of the fanning out process during its early stages can closely be approximated by a branching process which starts with a single progenitor, and in which the number of children of each individual, independently of the behavior of all other individuals has the distribution $BIN(n-1, c/n)$. In this branching process, let us say that an individual is *mortal* if his total number of descendants is finite. Let q_n be the probability that the progenitor is mortal and none of its descendants are of *type A*, given that the progenitor is not of *type A*. To determine the behavior of q_n as n tends to infinity, note that $BIN(n-1, c/n)$ converges in distribution to the Poisson distribution with mean c. Consider a *Poisson branching process* in which the number of children of any individual, independently of the behavior of other individuals has the Poisson distribution with mean c. Let q be the probability that the progenitor of this process is mortal and has no descendants of *type A*, given that the progenitor is not of *type A*. The conditional probability that an individual is mortal and has no descendants of *type A* given that it has k children and it is not of *Type A* is $(k_2 q)^k$. Unconditioning, we get $q = \sum_{k=0}^{\infty} e^{-c} \frac{c^k}{k!} (k_2 q)^k$, which leads to the equation $q = e^{-c} e^{ck_2 q}$. Substituting $\frac{1-x}{k_2}$ for q in the above we get $1 - x = k_2 e^{-cx}$. Hence, $q = \frac{1-\Theta'}{k_2}$.

□

Corollary 3.2 *$Pr[X(r)$ contains no Type A vertex $\mid$ $X(r)$ is small $]$ tends to $\frac{1-\Theta'}{1-\Theta}$.*

Proof: Note that $Pr[X(r)$ is small $]$ tends to $1 - \Theta$ and the corollary follows from Lemma 3.2.
□

Lemma 3.3 $Pr[\delta_u = 0 | \delta_v = 1] \geq Pr[\delta_u = 0] - O(\log^2 n/n)$

Proof: A proof similar to Lemma 3.1 will work for this lemma. □

Corollary 3.3 $Pr[\delta_u = 1 | \delta_v = 1] \leq Pr[\delta_u = 1] + O(\log^2 n/n)$

Theorem 3.5 *Let* $S = |\{u \mid \delta_u = 1\}|$. *Then, with probability tending to 1,* $|S - (1 - \Theta')n| < w(n)\sqrt{n \log^2 n}$.

Proof: Since, $S = \sum_{r=1}^{n} \delta_r$, $E(S) = (1 - \Theta')n$. Also, $E(S^2) = E(S) + n(n-1)E(\delta_u \delta_v)$, where u and v are any two distinct vertices. By Lemma 3.3 we have that $E(\delta_u \delta_v) \leq E(\delta_u)(E(\delta_v) + O(\log^2 n/n))$. It now follows that variance of S is $O(n \log^2 n)$ and the desired conclusion now follows from Chebyshev's inequality. □

We will need the following two refinements of the results of Section 2.2 for our future discussions.

Lemma 3.4 *With probability tending to 1,* $|X(r)|$ *(recall Definition 2.1) in a digraph* D *drawn from* $D_{n,c/n}$ *will be bounded above by* $\ln \ln n$, *given that* $X(r)$ *is small.*

Proof: We will need Markov's inequality, which states that for a random variable X that takes only nonnegative values, $Pr[X \geq \beta] \leq \frac{E(X)}{\beta}$, where $\beta > 0$ and $E(X)$ is the expected value of X. The result now follows from Markov's inequality and the fact that given $X(r)$ is small, the expected size of $|X(r)|$ is a constant (Theorem 2.2). □

Lemma 3.5 *Recall that an exceptional vertex is one that lies in a strong component of size greater than* 1 *but not in the unique largest strong component. Let the random variable* s *be the number of exceptional vertices in a digraph* D *drawn from* $D_{n,c/n}$. *Then with probability tending to* 1, s *is bounded above by* $\ln \ln n$.

Proof: The result follows from Markov's inequality and the fact that the expected size of s is a constant (Theorem 2.5). □

4 Main Results

In this section, by size of the answer to a program, we mean the number of tuples in the fixpoint of the recursively defined relation in that program. We recall that the a and the b graphs are drawn from $D_{n,c/n}$ and Θ will denote the unique root in $[0, 1]$ of the equation $1 - x - e^{-cx} = 0$. We will use the graph theoretic formulation for the problem of finding the sizes of the fixpoints of the recursively defined relations we derived in Section 2.1 to derive our bounds.

Recall that *LARGEOUT* is the set $\{u | X(u)$ is large$\}$, *LARGEIN* is the set $\{v | Y(v)$ is large$\}$, and *LARGE* is the set $\{u | X(u)$ and $Y(u)$ is large$\}$. We will denote *LARGEOUT* of the a graph by $LOUT_a$ and *LARGEIN* of the a graph by LIN_a. Similarly, we will denote *LARGEOUT* of the b graph by $LOUT_b$ and *LARGEIN* of the b graph by LIN_b.

4.1 Transitive closure

Here, we derive analytic bounds for the sizes of answers to the TC, TC_{mg}, and TC_{factor} Datalog programs.

Lemma 4.1 *Let D be drawn from $D_{n,c/n}$. Let x be some vertex in* LARGEOUT *of D. With probability tending to 1, the number of vertices not in* LARGEIN *that are reachable from x is at most $O(n^\epsilon), 0 < \epsilon < 1$.*

Proof: If x belongs to *LARGEIN* also, the lemma is trivially true. If not, let y be some vertex in *LARGE*. From Theorem 2.1, the forward reachability of y is at least $\Theta n - w(n)\sqrt{n}$ and the forward reachability of x is at most $\Theta n + w(n)\sqrt{n}$. Since y is in the forward reachability of x, it follows that the number of vertices not in *LARGEIN* that are reachable from x is at most $2w(n)\sqrt{n}$, which is $O(n^\epsilon)$. □

Theorem 4.1 *With probability tending to 1, the size of the answer to TC is $\Theta^2 n^2 + O(n^{1+\epsilon}), 0 < \epsilon < 1$.*

Proof: Let $x \in LOUT_a$. Then every vertex $y \in LIN_a$ is reachable from x and hence in the answer to TC. It now follows from Theorem 2.3 that the number of vertex pairs (x, y) such that $x \in LOUT_a$ and $y \in LIN_a$ that will be in the answer to TC is $\Theta^2 n^2 + O(n^{1+\epsilon})$. By Lemma 4.1, the number of vertices $y \notin LIN_a$ that are reachable from x is $O(n^\epsilon)$. Therefore, the number of vertex pairs (x, y) such that $x \in LOUT_a$ and $y \notin LIN_a$ that will be in the answer to TC is $O(n^{1+\epsilon})$.

Let $x \notin LOUT_a$. By Theorem 2.1, the forward reachability of x is at most $B \ln n$, for some constant B. This means the number of vertex pairs (x, y) such that $x \notin LOUT_a$ that are in the answer to TC is $O(n^{1+\epsilon})$. □

Theorem 4.2 *With probability tending to 1, the size of the answer to TC_{mg}, if the magic set is big, is $\Theta^3 n^2 + O(n^{1+\epsilon})$, $0 < \epsilon < 1$.*

Proof: Recall that the magic set is the reachability set from vertex 1 in the a graph. If the magic set is big, it contains all vertices of LIN_a and probably some vertices of $LOUT_a$ that are not in LIN_a. Let $x \in LARGE$ of the a graph. Then every vertex $y \in LIN_a$ is reachable from x and hence in the answer to TC_{mg}. Moreover no other vertex is reachable from x. It now follows from Theorem 2.3 that the number of vertex pairs (x, y) such that $x \in LARGE$ that will be in the answer to TC is $\Theta^3 n^2 + O(n^{1+\epsilon})$.

Let x be in $LOUT_a$ but not in LIN_a. By Lemma 4.1, the number of vertices $x \notin LIN_a$ that are reachable from 1 is $O(n^\epsilon)$. Therefore, the number of vertex pairs (x, y) such that x is in $LOUT_a$ but not in LIN_a that will be in the answer to TC_{mg} is $O(n^{1+\epsilon})$.

Let $x \notin LOUT_a$. By Theorem 2.1, the forward reachability of x is at most $B \ln n$, for some constant B. This means the number of vertex pairs (x, y) such that $x \notin LOUT_a$ that are in the answer to TC_{mg} is $O(n^{1+\epsilon})$. □

Theorem 4.3 *The expected size of the answer to TC_{mg} tends to $\Theta^4 n^2 + O(n^{1+\epsilon})$, $0 < \epsilon < 1$.*

Proof: The result follows from the fact that the probability that the magic set is big is Θ and an upper bound on the size of the answer to TC_{mg} when the magic set is small is $O(n^{1+\epsilon})$. □

Theorem 4.4 *With probability tending to 1, the expected size of the answer to TC_{factor} is $\Theta^2 n + O(n^{\epsilon})$, $0 < \epsilon < 1$.*

Proof: The answer to TC_{factor} consists of all tuples reachable from 1. If the magic set is big, the size of the answer is $\Theta n + O(n^{\epsilon})$. If the magic set is small, the size of the answer is $O(n^{\epsilon})$. The result now follows from the fact that the probability that the magic set is big is Θ. □

4.2 Same Generation

Here, we derive analytic bounds for the sizes of answers to the SG, SG_{mg}, and SG_{factor} Datalog programs.

The following theorem identifies a set of vertices that will almost certainly be in the answer to SG. That gives a lower bound on the size of the materialized relation for SG.

Theorem 4.5 *With probability tending to 1, the size of the answer to SG will be at least $\Theta^2 n^2 + O(n^{1+\epsilon}), 0 < \epsilon < 1$.*

Proof: We know that there will be more than $\ln n$ vertices in LIN_a with probability tending to 1. It follows from Theorem 3.4 that the probability that all these $\ln n$ vertices will have a small reverse reachability in the b graph tends to zero. Therefore, with probability tending to 1, there exists at least one vertex which is in LIN_a as well as in LIN_b. Let z_0 be one such vertex.

Let k_1 and k_2 be two natural numbers such that $gcd(k_1, k_2) = 1$. Then it is a well known fact that every natural number $N > N_0$, for some N_0, can be expressed as $ik_1 + jk_2$ where i and j are some non-negative integers. We can then conclude, from the above and Theorem 3.1, that there exists a positive integer N_a such that for $N > N_a$, every vertex x in $LOUT_a$, will have a path of length N to z_0 in the a graph. Similarly, there exists a positive integer N_b such that every vertex $y \in LOUT_b$ will have a path of length N to z_0 in the b graph, where $N > N_b$. This proves that every vertex pair (x, y) such that $x \in LOUT_a$ and $y \in LOUT_b$ will be in the answer to SG. The number of such vertex pairs is $\Theta^2 n^2 + O(n^{1+\epsilon})$ from Theorem 2.3 and the desired result follows. □

We will now prove that the lower bound of the previous theorem is in fact an upper bound too. In other words, the contribution from the vertex pairs other than those identified in the previous theorem is of a smaller order than n^2.

The following definitions are needed before we can proceed to the next theorem.

Definition 4.1 The set $R(i)$ is defined as follows for $i = 1, \ldots, n$: $R(i)$ is the set of vertex pairs (x, y) such that the following are true

- $x \notin LOUT_a$ or $y \notin LOUT_b$
- There exists some non negative integer k such that there is a directed path of length k from x to i in the a graph and a directed path of length k from y to i in the b graph.

□

Definition 4.2 The set $P(i)$ is defined as follows:

$$P(i) = \{(x, y) \mid (x, y) \in R(i) \wedge x \notin LOUT_a\}$$

and the set $Q(i)$ is defined as follows:

$$Q(i) = \{(x, y) \mid (x, y) \in R(i) \wedge x \in LOUT_a\}$$

□

Theorem 4.6 *With probability tending to* 1 *, an upper bound on the size of the answer to SG will be* $\Theta^2 n^2 + O(n^{1+\epsilon})$, $0 < \epsilon < 1$.

Proof: We have shown in the last theorem that with probability tending to 1, all vertex pairs (x, y) where $x \in LOUT_a$ and $y \in LOUT_b$ will be in the answer to SG. We will now derive an upper bound on the number of vertex pairs (x, y), where $x \notin LOUT_a$ or $y \notin LOUT_b$, that will be in the answer to SG. We will show that the number of such vertex pairs is $O(n^{1+\epsilon})$, $0 < \epsilon < 1$.

It can be seen that $\sum_{i=1}^{n} R(i)$ is an upper bound on the number of distinct vertex pairs (x, y) such that $x \notin LOUT_a$ or $y \notin LOUT_b$ that are produced as answer to SG . We will derive an upper bound on $R(r)$ and then multiply it by n to get an upper bound on the above sum. We consider a case by case analysis based on the position of r in the a and b graphs.

CASE 1 : $r \in LOUT_a$ and $r \in LOUT_b$
$R(r) = 0$ from the definition of $R(r)$.

CASE 2 : $r \notin LOUT_a$ and $r \notin LOUT_b$
Since $R(r) = P(r) \cup Q(r)$, we will derive upper bounds on $P(r)$ and $Q(r)$ and add them up to get an upper bound on $R(r)$. The argument for both cases is similar, we will present the argument for $P(r)$ alone.

Let $S = \{x \mid (x, y) \in P(r)\}$. All vertices in S have a forward reachability which is at most $\ln \ln n$ by Lemma 3.4. It is now easy to see that $|S|$ is less than or equal to $Z(r)$, where $Z(r)$ is the number of vertices that have a path to r of length $\ln \ln n$ or less in the a graph. We can

therefore infer from Theorem 3.2 that $|S|$ is $O(n^{1/m})$. If $Y(r)$ (recall Definition 2.1) of the a graph does not have an exceptional vertex in it, all paths from vertices in S to r in the a graph are simple paths (a path where no vertex is repeated). If they were not simple, then there is a directed cycle and that implies the presence of a strong component of size greater than 1. Therefore *all* paths from the vertices in S to r in the a graph are at most $\ln \ln n$ in length, since a simple path of length greater than $\ln \ln n$ would imply a reachability bigger than $\ln \ln n$. The number of y's that can now be paired with vertices in S is at most $O(n^{1/m})$, from Theorem 3.2. The reason here is that the length of the path from y to r in the b graph has to be the same as the length of the path from x to r in the a graph, if x and y are to be paired. So, if $Y(r)$ of the a graph does not contain an exceptional vertex, an upper bound on $P(r)$ is $O(n^{2/m})$. If $Y(r)$ of the a graph has an exceptional vertex in it, the vertices in S could potentially have long paths to r in the a graph and at worst $O(n)$ y's could be paired with x and so an upper bound on $P(r)$ in this case is $O(n^{1+1/m})$.

CASE 3 : r is in $LOUT_a$ and r is not in $LOUT_b$
This case is argued similar to Case 2 and we conclude that $R(r)$ is $O(n^{2/m})$ whenever $Y(r)$ in the b graph does not contain an exceptional vertex. Otherwise $R(r)$ is $O(n^{1+1/m})$.

CASE 4 : r is not in $LOUT_a$ and r is in $LOUT_b$
This case is argued similar to Case 2 and we conclude that $R(r)$ is $O(n^{2/m})$ whenever $Y(r)$ in the a graph does not contain an exceptional vertex. Otherwise $R(r)$ is $O(n^{1+1/m})$.

The number of vertices in the a graph that have an exceptional vertex in their reverse reachability set is at most $O((\ln \ln n)^2)$, since there are at most $\ln \ln n$ exceptional vertices (Lemma 3.5) in the a graph and each of these vertices has a forward reachability of at most $\ln \ln n$. Similarly, the number of vertices in the b graph that have an exceptional vertex in their reverse reachability set is at most $O((\ln \ln n)^2)$. Therefore, the number of vertices that have an exceptional vertex in their reverse reachability set in at least one of the a or b graphs is at most $O((\ln \ln n)^2)$.

Therefore, $\sum_{i=1}^{n} R(i)$ is $O(n^{1+1/m}(\ln \ln n)^2 + n^{1+2/m})$. The first term is the contribution of those i's that have an exceptional vertex in $Y(i)$ of the a graph or in $Y(i)$ of the b graph and the second term is the contribution of those i's that have no exceptional vertices in either of their reverse reachability sets. It follows now that the sum is $O(n^{1+\epsilon})$, for $0 < \epsilon < 1$. □

Theorem 4.7 *With probability tending to* 1, *the size of the answer to* SG_{mg}, *if the magic set is big, is at least* $\Theta^3 n^2 + O(n^{1+\epsilon})$, $0 < \epsilon < 1$.

Proof: Recall that the magic set is the reachability set from vertex 1 in the a graph. If the magic set is big, then all the vertices that are in LIN_a are in the magic set. We can now prove similar to the proof for SG that every vertex pair (x, y) such that $x \in LARGE$ of the a graph and $y \in LOUT_b$ will be in the answer to SG_{mg}. Therefore, from Theorem 2.3, the number of such vertex pairs is $\Theta^3 n^2 + O(n^{1+\epsilon})$ and the desired result for SG_{mg} follows. □

Theorem 4.8 *With probability tending to* 1, *an upper bound on the size of the answer to* SG_{mg}, *if the magic set is big, is* $\Theta^3 n^2 + O(n^{1+\epsilon})$, $0 < \epsilon < 1$.

Proof: We have seen in the previous theorem that, if the magic set is big, then all the vertices of LIN_a are in the magic set and all vertex pairs (x, y) such that $x \in LARGE$ of the a graph and $y \in LOUT_b$ will be in the answer to SG_{mg}. If 1 is not in $LARGE$ of the a graph, then the vertices of the magic set that are in $LOUT_a$ but not in LIN_a form a subset of the vertices that have a path from 1 of length less than $\ln \ln n$. From Theorem 3.3 we can conclude that the number of vertices in the Magic set that are in $LOUT_a$ but not in LIN_a is $O(n^{1/m})$ and the contribution to the answer is going to be at most $O(n^{1+1/m})$ from these extra vertices. We are left to deal with the contribution from vertex pairs (x, y) such that $x \notin LOUT_a$. A reasoning similar to the proof of Theorem 4.6 shows that this contribution is $O(n^{1+\epsilon})$. □

Theorem 4.9 *The expected size of the answer to* SG_{mg} *tends to* $\Theta^4 n^2 + O(n^{1+\epsilon})$, $0 < \epsilon < 1$.

Proof: The proof follows from the fact that the probability that the magic set is big is Θ and an upper bound on the size of the answer to SG_{mg} when the magic set is small is $O(n^{1+\epsilon})$. □

4.3 Canonical Factorable Recursion

Here, we derive analytic bounds for the sizes of answers to the CFR, CFR_{mg}, and CFR_{factor} Datalog programs.

Theorem 4.10 *Let* Θ' *be the unique root in* $[0, 1]$ *of the equation* $1 - x = (1 - \Theta)e^{-cx}$. *With probability tending to* 1, *the size of the answer to* CFR *is* $\Theta(2\Theta' - \Theta)n^2 + O(n^{1+\epsilon})$,$0 < \epsilon < 1$

Proof: Recall that (x, y) is in the answer if and only if there is a path from x to some vertex z_0 in the a graph and a path from y to z_0 in the b graph. For a fixed i the number of vertices, that do not belong to $LOUT_a$ and have a path to i is less than or equal to $Z(i)$, and hence, $O(n^{1/m})$ by Theorem 3.2. Similarly, the number of vertices that do not belong to $LOUT_b$ and have a path to i in the b graph is $O(n^{1/m})$. Therefore, for a fixed i, there are at most $O(n^{2/m})$ vertex pairs (x, y) such that $x \notin LOUT_a$ and $y \notin LOUT_b$ and x has a path to i in the a graph and y has a path to i in the b graph. Hence, the number of vertex pairs (x, y) such that $x \notin LOUT_a$ and $y \notin LOUT_b$ that are produced as an answer to CFR is bounded above by $O(n^{1+\epsilon})$.

Consider the vertex pairs (x, y) such that $y \in LOUT_b$. Let us partition the vertex set of the a graph as follows: Mark the vertices that belong to LIN_b as *Type A* in the a graph and the rest as *Type B*. Let δ_r be a $0-1$ random variable as defined in Definition 3.1. Recall that $\delta_r = 1$ if $X(r)$ is small and $X(r)$ contains no *Type A* vertex, and $\delta_r = 0$ otherwise. It can be seen that $k_1 = \Theta$ (the number of vertices of *Type A*) and $k_2 = 1 - \Theta$. For a fixed r, if $\delta_r = 1$, $X(r)$ (in the a graph) contains only vertices that are not in LIN_b in the b graph. So, each vertex of $X(r)$ will have a reverse reachability in the b graph of $O(\ln n)$. Further, $|X(r)|$ is $O(\ln n)$. Therefore, the number of vertex pairs of the form (r, y) in the answer is $O(\ln^2 n)$. Therefore, there will be

at most $O(n \ln^2 n)$ vertex pairs of the form (x, y), where $y \in LOUT_b$. For a fixed x, if $\delta_x = 0$, then all vertex pairs (x, y) such that $y \in LOUT_b$ will be in the answer. The reason is that if $\delta_x = 1$, then it either has a large reachability, in which case a proof similar to Theorem 4.5 will prove the preceding claim; or it has a small reachability that contains a *Type A* vertex. Therefore, total number of vertex pairs (x, y) such that $\delta_x = 0$ and $y \in LOUT_b$ that will be in the answer is $(n - \sum_{x=1}^{n} \delta_x)(\Theta n)$. This expression is $\Theta'\Theta n^2 + O(n^{1+\epsilon})$ by Theorem 3.5.

Consider the vertex pairs (x, y) such that $x \in LOUT_a$. This will also yield $\Theta'\Theta n^2 + O(n^{1+\epsilon})$ by a similar argument as above. But, we have counted the vertex pairs (x, y) such that $x \in LOUT_a$ and $y \in LOUT_b$ twice. So subtracting $\Theta^2 n^2$ for that, we get the final expression $\Theta(2\Theta' - \Theta)n^2 + O(n^{1+\epsilon})$. □

Theorem 4.11 *Let Θ' be the unique root in $[0, 1]$ of the equation $1 - x = (1 - \Theta)e^{-cx}$. With probability tending to 1,*

1. *The size of the answer to CFR_{mg}, when the magic set is big, is $\Theta^2(2\Theta' - \Theta)n^2 + O(n^{1+\epsilon})$, $0 < \epsilon < 1$*
2. *The size of the answer to CFR_{factor} when the magic set is big is $\Theta' n + O(n^{\epsilon})$ and the expected size of the answer to CFR_{factor} when the magic set is small is $(1 - \frac{1-\Theta'}{1-\Theta})\Theta n + O(n^{\epsilon})$, $0 < \epsilon < 1$.*

Proof: Notice that the expression for CFR_{mg} is exactly Θ times the expression for CFR. The extra factor of Θ comes about because only vertex pairs (x, y) where x belongs to magic set can appear in the answer to CFR_{mg}. An argument similar to the proof of Theorem 4.10 will give the desired result for CFR_{mg}.

Recall that y belongs to the answer to CFR_{factor} if and only if there is a path from y to z_0 in the b graph, where z_0 belongs to the magic set. We now analyze CFR_{factor} when the magic set is big. Let us now partition the vertex set of the b graph as follows: Mark the vertices that belong to the magic set as *Type A* and the rest as *Type B*. Let δ_r be a $0 - 1$ random variable as defined in Definition 3.1. It can be seen that $k_1 = \Theta$ and $k_2 = 1 - \Theta$. All vertices r such that $\delta_r = 0$ will be in the answer (and no other vertex will be in the answer). This is because from previous arguments, it follows that all the vertices in $LOUT_b$ will be in the answer and all vertices of the b graph that have a vertex of the magic set in $X(r)$ will also be in the answer. Therefore, the desired result for CFR_{factor}, when the magic set is big, follows from Theorem 3.5.

We now turn to CFR_{factor} when magic set is small. Notice that what we actually need to know in this case is if any of the vertices in the magic set is in *LARGEIN* of the b graph. If not, then we will have $O(n^{\epsilon})$ vertices only in the answer. If there is at least one, then we will have $\Theta n + O(n^{\epsilon})$ vertices in the answer. It then follows from Corollary 3.2 that the probability of the second event is $1 - \frac{1-\Theta'}{1-\Theta}$. Hence the desired result follows when magic set is small for CFR_{factor}. □

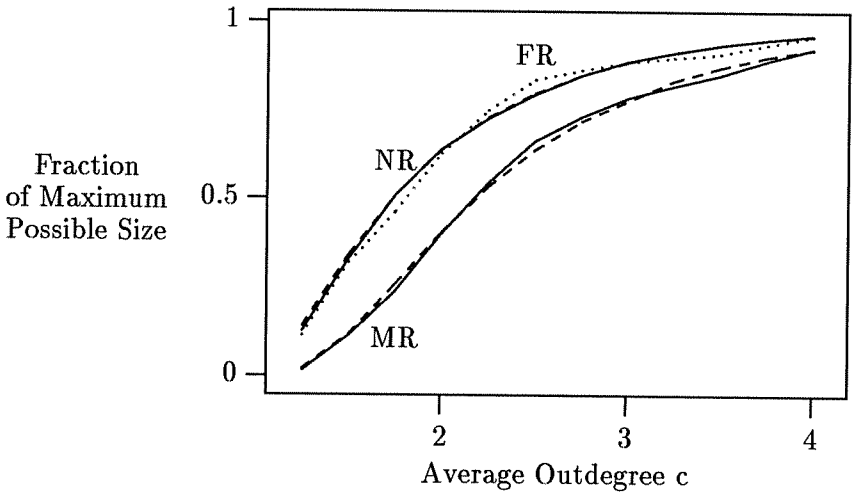

Figure 2: Results for TC Programs

Theorem 4.12 *Let Θ' be the unique root in $[0,1]$ of the equation $1-x=(1-\Theta)e^{-cx}$.*

1. *The expected size of the answer to CFR_{mg} tends to $\Theta^3(2\Theta'-\Theta)n^2+O(n^{1+\epsilon})$, $0<\epsilon<1$*
2. *The expected size of the answer to CFR_{factor} tends to $\Theta'\Theta n+(1-\Theta)(1-\frac{1-\Theta'}{1-\Theta})\Theta n+O(n^{\epsilon})$, which simplifies to $\Theta(2\Theta'-\Theta)n+O(n^{\epsilon})$, $0<\epsilon<1$.*

Proof: If the magic set is small, the size of the answer to CFR_{mg} is $O(n^{1+\epsilon})$. Since, the probability the magic set is big is Θ, the expression for the expected size of the answer to CFR_{mg} follows from Theorem 4.11.

The result for CFR_{factor} follows from Theorem 4.11 and the fact that the probability the magic set is big is Θ. □

5 Experimental Results

In this section we describe experimental results that demonstrate the convergence of the above results. We generated random digraphs for various values of n and p and we wrote simple programs to compute the tuples that would have been generated by a fixpoint evaluation algorithm with those random graphs as the base relations.

We present one set of experiments here wherein we held the number of nodes in the graph constant at 256, while we varied c from 1.25 to 4.0, in steps of 0.25. The results for the transitive closure program is shown in Figure 2, the results for the same generation program is shown in Figure 3, and the results for the canonical factorable program is shown in Figure 4. The dashed line in all cases is the coefficient of the highest order term in the corresponding analytic expression. The lower solid line (labeled *MR*) in each figure gives the average number of tuples materialized by Magic Sets rewriting strategy, divided by n^2. The upper solid line (labeled

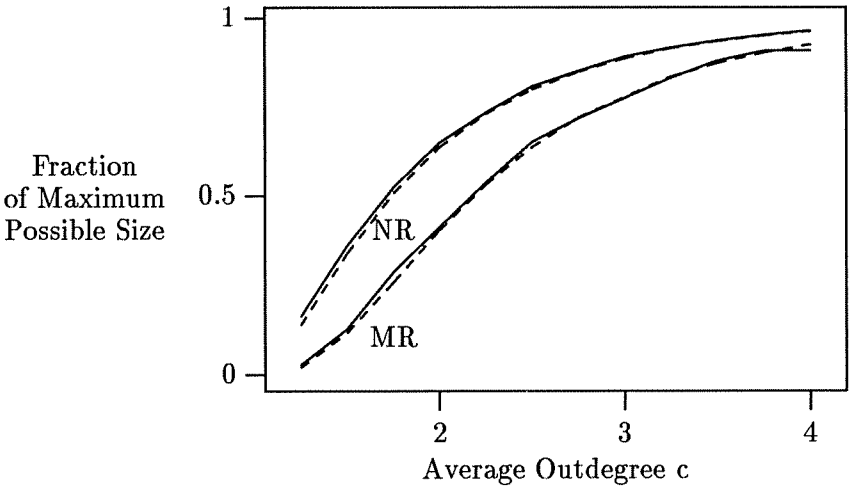

Figure 3: Results for Same Generation Programs

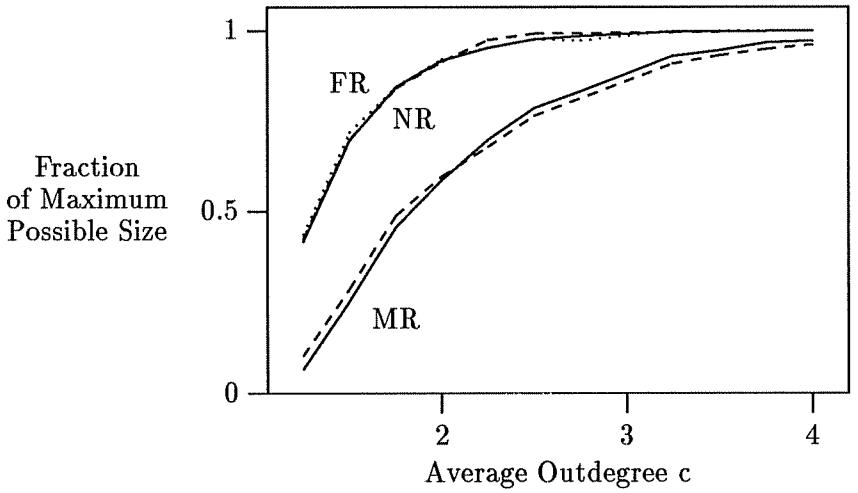

Figure 4: Results for CFR Programs

NR) in each figure gives the average number of tuples materialized by the program without any rewriting, divided by n^2. The dotted line (labeled *FR*) in Figures 2 and 4 gives the average number of tuples materialized by the Factoring rewriting algorithm, divided by n (as opposed to n^2 for other cases). It can be seen that in all cases, the results are in close agreement with the analytic formulas.

6 Conclusion

We derived analytic formulas for the sizes of materialized relations for the "same generation" and the "canonical factorable recursion". We considered the programs without rewriting as well as the rewritten programs produced by the Magic Sets Strategy and Factoring technique in response to selection queries. We ran experiments that agree closely with the analytic formulas. Our results demonstrate that unless the arity of the recursive predicate can be reduced or the base relations are extremely sparse, pushing selections by restricting the fixpoint evaluation produces only a small savings over evaluating the original, non-rewritten program.

The approach we adopted was to use random digraphs to model the base relations. Then, the question of the size of the materialized relation transforms into a question about the properties of a random graph or the interaction between some number of random graphs. We believe that this approach can be used effectively for a wider class of recursions than that considered here; furthermore, we conjecture that most Datalog recursions will exhibit the key property that as the "density" of the base relations increases above some low threshold, the size of the recursively defined relation will quickly approach its worst-case bound, and virtually all tuples in this relation will be relevant to the computation of selection queries on this relation.

Acknowledgements

We would like to thank G. Ramalingam for providing useful comments.

References

[AKS81] M. Ajtai, J. Komlós, and E. Szemerédi. The longest path in a random graph. *Combinatoria*, 1, pages 1–12, 1981.

[BKBR87] Catriel Beeri, Paris Kanellakis, Francois Bancilhon, and Raghu Ramakrishnan. Bounds on the propagation of selection into logic programs. In *Proceedings of the ACM Symposium on Principles of Database Systems*, pages 214–226, San Diego, California, March 1987.

[BMSU86] Francois Bancilhon, David Maier, Yehoshua Sagiv, and Jeffrey D. Ullman. Magic sets and other strange ways to implement logic programs. In *Proceedings of the ACM Symposium on Principles of Database Systems*, pages 1–15, Cambridge, Massachusetts, March 1986.

[Bol85] Béla Bollobás. *Random Graphs.* Academic Press, London, 1985.

[BR87] Catriel Beeri and Raghu Ramakrishnan. On the power of magic. In *Proceedings of the ACM Symposium on Principles of Database Systems*, pages 269–283, San Diego, California, March 1987.

[BR88] Francois Bancilhon and Raghu Ramakrishnan. Performance evaluation of data intensive logic programs. In Jack Minker, editor, *Foundations of Deductive Databases and Logic Programming*, pages 439–517, Los Altos, California, 1988. Morgan Kaufmann.

[GKS91] S. Ganguly, R. Krishnamurthy, and A. Silberschatz. An analysis technique for transitive closure algorithms: a statistical approach. In *Proceedings of the IEEE Data Engineering Conference*, 1991. To appear.

[HL86] Jiawei Han and Hongjun Lu. Some performance results on recursive query processing in relational database systems. In *Proceedings of the International Conference on Data Engineering*, pages 533–541, 1986.

[HN88] Ramsey W. Haddad and Jeffrey F. Naughton. Counting methods for cyclic relations. In *Proceedings of the ACM Symposium on Principles of Database Systems*, pages 333–340, Austin, Texas, March 1988.

[Kar90] Richard M. Karp. The transitive closure of a random digraph. *Random Structures and Algorithms*, 1(1):73–93, 1990.

[MSPS87] Alberto Marchetti-Spaccamela, Antonella Pelaggi, and Domenico Sacca. Worst-case complexity analysis of methods for logic query implementation. In *Proceedings of the ACM Symposium on Principles of Database Systems*, pages 294–301, San Diego, California, March 1987.

[Nau88] Jeffrey F. Naughton. Benchmarking multi-rule recursion evaluation strategies. Technical Report CS-TR-141-88, Princeton University, 1988.

[NRSU89] Jeffrey F. Naughton, Raghu Ramakrishnan, Yehoshua Sagiv, and Jeffrey D. Ullman. Argument reduction through factoring. In *Proceedings of the Fifteenth International Conference on Very Large Databases*, pages 173–182, Amsterdam, The Netherlands, August 1989.

[Rag86] P. Raghavan. Probabilistic construction of deterministic algorithms: Approximating packing integer programs. In *Proceedings of the 27th Annual IEEE Symposium on Foundations of Computer Science*, pages 10–18, 1986.

[Ram88] Raghu Ramakrishnan. Magic templates: A spellbinding approach to logic programs. In *Proceedings of the International Conference on Logic Programming*, pages 140–159, Seattle, Washington, August 1988.

[RS62] J. B. Rosser and L. Schoenfeld. Approximate formulas for some functions of prime numbers. *Illinois Journal of Mathematics*, 6:64–94, 1962.

[SZ87] Domenico Sacca and Carlo Zaniolo. Magic counting methods. In *Proceedings of the ACM-SIGMOD Symposium on the Management of Data*, pages 49–59, San Fransisco, California, June 1987.