On Sums of Independent Random Variables with Unbounded Variance, and Estimating the Average Degree in a Graph

Uriel Feige
Department of Computer Science and Applied Mathematics
The Weizmann Institute
Rehovot 76100, Israel
uriel.feige@weizmann.ac.il

ABSTRACT

We prove the following inequality: for every positive integer n and every collection $X_1, \ldots, X_n$ of nonnegative independent random variables that each has expectation 1, the probability that their sum remains below $n+1$ is at least $\alpha > 0$. Our proof produces a value of $\alpha = 1/13 \simeq 0.077$, but we conjecture that the inequality also holds with $\alpha = 1/e \simeq 0.368$.

As an example for the use of the new inequality, we consider the problem of estimating the average degree of a graph by querying the degrees of some of its vertices. We show the following threshold behavior: approximation factors above 2 require far less queries than approximation factors below 2. The new inequality is used in order to get tight (up to multiplicative constant factors) relations between the number of queries and the quality of the approximation. We show how the degree approximation algorithm can be used in order to quickly find those edges in a network that belong to many shortest paths.

Categories and Subject Descriptors

G.3 [**Mathematics of Computing**]: Probability and Statistics; F.2.2 [**Theory of Computing**]: Analysis of Algorithms and Problem Complexity—*Nonnumerical Algorithms and Problems*

General Terms

Theory, Algorithms

Keywords

inequalities, shortest paths

1. A NEW INEQUALITY

For a random variable X, its typical value may be very different from its mean. In particular, the probability that X exceeds its mean may be arbitrarily close to 1. In some special cases (e.g., when X is symmetric around its mean), the probability that X exceeds its mean is at most $1/2$. The purpose of this manuscript is to investigate the probability that X exceeds its mean when X is the sum of n independent random variables. We show that for nonnegative random variables, this probability is bounded away from 1, provided that we give ourselves a little slackness in exceeding the mean.

THEOREM 1. *Let $X_1, \ldots, X_n$ be arbitrary nonnegative independent random variables, with expectations $\mu_1, \ldots, \mu_n$ respectively, where $\mu_i \leq 1$ for every i. Let $X = \sum_{i=1}^{n} X_i$, and let μ denote the expectation of X (hence, $\mu = \sum_{i=1}^{n} \mu_i$). Then for every $\delta > 0$,*

$$Pr[X < \mu + \delta] \geq \min[\delta/(1+\delta), 1/13] \qquad (1)$$

The term $\delta/(1+\delta)$ in Theorem 1 is best possible, as one can take $X_1 = 1+\delta$ with probability $1/(1+\delta)$ and 0 otherwise, and all of the other X_i as the constant 1. This gives $\mu_i = 1$ for every i. For this case $Pr[X < \mu + \delta] = Pr[X_1 = 0] = \delta/(1+\delta)$. For large δ (e.g., $\delta = 1$), it is not true that $Pr[X \leq \mu + \delta] \geq \delta/(1+\delta)$. One can take for every i, $X_i = n + \delta$ with probability $1/(n+\delta)$ and 0 otherwise. This gives $\mu_i = 1$ for every i, implying $\mu = n$. For this case $Pr[X < n + \delta] = (1 - 1/(n+\delta))^n$, which is roughly $1/e$ for large n.

It is our conjecture that for every value of δ and n, one of the two examples above is the worst case for $Pr[X < \mu + \delta]$. The conjecture, if true, would allow us to replace the constant $1/13$ by $1/e$ in Theorem 1.

In may be instructive to consider how some standard probabilistic tools relate to Theorem 1. Consider the case that the X_i are identically distributed. Then the central limit theorem implies that when n is large enough, X approaches the normal distribution and hence $Pr[X < \mu]$ approaches $1/2$. However, in our Theorem 1 the variables X_i may depend on n, and hence n cannot be thought of as being "large enough" with respect to the X_i (even if they are i.i.d.). This relates to the fact that we place no bounds on the variance of the X_i, and hence standard bounds on deviations of random variables from their expectation (such as Chebyschev's bound, or Chernoff's bound) are not applicable. The only restriction on the random variables (other than being independent) is their nonnegativity. In particular, this means that X is nonnegative, and that Markov's inequality can

STOC'04, June 13–15, 2004, Chicago, Illinois, USA.

be used to show that $Pr[X \leq \mu + \delta] \geq \delta/(\mu + \delta)$. For the sum of independent identically distributed random variables, this bound tends to 0 as n grows (unlike the bound in Theorem 1).

The author is aware of some work of nature similar to Theorem 1. There are results surveyed and developed by Siegel [5] that show that under certain conditions the median of the sum of random variables does not exceed the mean. This holds for example for the sum of Bernoulli random variables (if the mean is an integer). The book "How to gamble if you must" by Dubins and Savage [3] analyses strategies for gambling when the goal is to maximize the probability of ending up with a profit of δ. There the strategies are adaptive (next gamble may depend on outcomes of previous gambles) and the gambler may quit once a net profit of δ is achieved. One of the main findings of [3] is a set of sufficient conditions under which the strategy of "playing boldly" is optimal. Informally, this strategy tries to reach a net profit of δ (taking into account also previous losses) in one gamble. A typical example is the repeated doubling approach to gain one dollar when there are 50/50 odds, in which the gambler first gambles one dollar, and then doubles the gamble until the first win (or until he/she runs out of money). The scenario in Theorem 1 can be viewed as a version of "how to gamble in parallel", in which n gambles with independent outcomes are placed in parallel in an attempt to reach a net profit of δ. Also in this case it seems that the best strategies (the author can think of) are based on hoping for one succesful gamble. Despite similarities in the nature of the results, the proof techniques from [3] and [5] do not appear to be applicable to the setting of Theorem 1.

Theorem 1 can in principle be used whenever one is interested in bounding the probability that the sum of independent random variables significantly exceeds its expectation. However, in many contexts the random variables are known to have some additional properties (e.g., bounded variance), and useful results can also be derived by other means. The application that motivated the development of the inequality (1) is described in Section 2.

2. ESTIMATING THE AVERAGE DEGREE

Let $G(V, E)$ be a graph with n vertices. A *degree query* specifies a vertex $v \in V$, and gets in reply d_v, the degree of v in G. We are interesting in estimating $m = |E|$ by making only degree queries. Equivalently, we wish to estimate the average degree $d = 2m/n$. We say that an algorithm provides a ρ estimation if its output d^* satisfies

$$d^* \leq d \leq \rho d^*.$$

Naturally, we limit our interest to $\rho \geq 1$. As our sampling based algorithms are randomized, there is some probability that their output fails to be a ρ estimation. We require this failure probability to be at most 1/3. We note that the failure probability can be reduced to an arbitrarily small value δ, by repeating the estimation algorithm $O(\log 1/\delta)$ times independently, and outputting the median of all estimates. Our goal is for given ρ, to design ρ estimation algorithms with as few queries as possible, and with failure probability at most 1/3.

Let us note here an observation that helps us to drastically reduce the number of queries in our algorithms. Consider first the case where rather than having an actual graph as input, the input is simply a sequence of integers $d_1, \ldots, d_n$, with the only restriction that for every i, $0 \leq d_i \leq n$. (For simplicity of the presentation we allow here values to range up to n, even though degrees can range only up to $n-1$.) Let $d = \frac{1}{n}\sum_{i=1}^{n} d_i$. We wish to estimate d. It is not hard to see that for any value d_0 (which one may think of as a large constant independent of n), $\Omega(n/d_0)$ queries are required in order to distinguish between the cases $d = 0$ and $d \geq d_0$. The reason is that it may happen that there are d_0 numbers with value n, and all other numbers have value 0. If we perform less than $n/2d_0$ queries, most likely we always get the 0 answer, which is exactly the answers that we would get if $d = 0$.

To get estimation algorithms with fewer queries, we shall use the fact that not every sequence $d_1, \ldots, d_n$ is a degree sequence of graphs. For example, if $d_1 = n - 1$, then necessarily $d_i \geq 1$ for all i. Still, the bad example given above can be modified to show that $O(n/d_0)$ queries are required in order to distinguish between the cases $d \leq d_0$ and $d \geq 2d_0 - O((d_0)^2/n)$. In the first of these two cases we can have all $d_i = d_0$. In the second of these two cases we can have $d_i = d_0$ for all vertices except for d_0 vertices of degree $n - 1$. Hence if we wish to have estimation algorithms with a sublinear (in n) number of queries, we need to restrict ourselves to $\rho \geq 2$.

There is one more restriction that we introduce. Observe that if G contains only one edge, one needs $\Omega(n)$ queries to distinguish this case from $d = 0$. To avoid the problem of handling such very sparse graphs (which are often not interesting anyway), we shall assume that $d \geq d_0$, for some d_0 that will be a parameter of our estimation algorithms. The reader may think of d_0 as typically having value at least 1. Hence the estimation algorithm is allowed to output $d^* = 0$ as an estimation of d for very sparse graphs, even though the ratio between d and d^* is in this case infinite.

As noted above, for $\rho < 2$ and $d_0 = 1$, the number of queries needed by an estimation algorithm might be $\Omega(n)$. Our main observation is that for $\rho > 2$ and for $d_0 = 1$, the number of queries in the estimation algorithm drops dramatically, from $\Omega(n)$ to $O(\sqrt{n})$. This result is stated in more technical terms in the following theorem.

THEOREM 2. *For $\epsilon > 0$, $\rho = 2 + \epsilon$, there is a ρ estimation algorithm for the average degree of a graph that uses $O(\frac{1}{\epsilon}\sqrt{n/d_0})$ queries, and is applicable to all graphs of average degree at least d_0.*

In terms of the application of estimating the average degree in the graph, the more interesting part of our upper bound on the number of queries is the term $\sqrt{n/d_0}$. The dependency on ϵ may be less interesting, especially if one is satisfied with large values of ϵ, such as $\epsilon = 1$. However, achieving a linear dependency in $1/\epsilon$ (rather than some polynomial dependency) is the part that uses Theorem 1.

In Section 4 we prove Theorem 2. In Section 5 we show how Theorem 2 can be used in order to obtain Theorem 3, addressing a problem that is studied in [2].

THEOREM 3. *There is a randomized algorithm that runs in time $O(\frac{mn \log n}{\epsilon\sqrt{c}})$ on graphs with n vertices and m edges, and outputs a list of edges that with high probability satisfies:*

1. *Every edge that is on at least c shortest paths is on the list.*
2. *No edge that is on less than $(1/2 - \epsilon)c$ shortest paths is on the list.*

3. PROOF OF THEOREM 1

In this section we prove Theorem 1.

PROOF. Fix n, δ, and arbitrary nonnegative random variables $X_1, \ldots, X_n$ with means at most 1. We prove that Equation (1) holds. We assume without loss of generality that the support of every random variable is composed of a finite set of values. (This is a standard argument, but we sketch it for completeness. Any value larger than $\mu+\delta$ in the support of a random variable can be lowered to $\mu+\delta$, without increasing the probability that $X < \mu+\delta$. Thereafter, any continuous random variable can be approximated by a discrete random variable with the same mean and whose support includes only multiples of ϵ, where ϵ is chosen to be much smaller than δ/n. For these new random variables, $X'_1, \ldots, X'_n$, the event $X' < \mu+\delta'$ where $\delta' = \delta - \epsilon n$ implies that for the original variables, $X < \mu+\delta$. By making ϵ arbitrarily small, we can make δ' arbitrarily close to δ.)

Our proof of Equation (1) consists of a sequence of transformations on the variables X_i. For simplicity of notation, we keep calling the random variables by X_i, their sum by X and the expectation of X by μ, even though the random variables themselves and μ do change by the transformations. The invariant kept by the transformations is that $Pr[X < \mu+\delta]$ does not increase (though the interpretation of X and μ, but not δ, does change). Other properties of the random variables may change by the transformations. In particular, the reduce support transformation (to be defined shortly) when applied to two random variables that were originally identically distributed might transform them to new random variables that are not identically distributed. Moreover, the merge transformation might generate random variables whose mean is larger than 1, even though all original random variables have mean at most 1.

Our first transformation, *remove constant*, is applied whenever there is a random variable X_i that is constant, that is, $Pr[X_i = \mu_i] = 1$. Such a random variable is removed, and μ is updated to $\mu - \mu_i$. Clearly, $Pr[X < \mu+\delta]$ does not change by *remove constant*.

The transformation *reduce support* is applied to every random variable whose support has at least three values, and replaces it with a new random variable with the same mean, and whose support includes at most two values from the original support.

LEMMA 1. *Let X_i be a random variable whose support includes at least three values. Then X_i can be replaced by a new variable (which we shall also call X_i) without changing μ_i, and whose support includes only two values from the original support of X_i. This can be done without increasing $Pr[X < \mu+\delta]$.*

PROOF. Let $\{v_1, \ldots, v_k\}$ be the support of the original X_i, and for $1 \le j \le k$, let q_j denote the conditional probability of the event $[X < \mu+\delta]$, conditioned on the event $[X_i = v_j]$. For the new X_i and for $1 \le j \le k$, we wish to select $p_j = Pr[X_i = v_j]$, under the restrictions that the mean of X_i is preserved, and that $Pr[X < \mu+\delta]$ does not increase. This can be expressed by the following linear program over the variables p_j:

Minimize $\sum_{j=1}^{k} q_j p_j$
subject to:

- $\sum_{j=1}^{k} p_j = 1$
- $\mu_i = \sum_{j=1}^{k} p_j v_j$
- $p_j \ge 0$, for every j.

The above linear program is feasible (as the probabilities associated with the original X_i satisfy the constraints). By the theory of linear programming, there is a *basic* optimal solution in which at most two p_j are nonzero. □

The transformation *align with 0* is applied to every random variable whose support has two values and these values are greater than 0 (say X_i has value v_1 with probability p and v_2 with probability $(1-p)$, with $0 < v_1 < v_2$), and replaces it by a random variable that has value $v_1 - v_1 = 0$ with probability p, and has value $v_2 - v_1$ with probability $(1-p)$. This decreases μ_i by v_1, and to compensate for it we update μ to $\mu - v_1$. Clearly, $Pr[X < \mu+\delta]$ does not change by *align with 0*.

The transformation *merge* takes the two random variables with smallest mean (say X_i and X_j), and replaces them by a new variable in three steps. First, replace X_i and X_j by a new random variable that is distributed like their sum $X_i + X_j$. Then apply *reduce support* to this new random variable. Finally, apply *align with 0* or *remove constant* to the new random variable (if applicable).

It is easy to see that the transformation *merge* does not increase $Pr[X < \mu+\delta]$.

The sequence of transformations that we perform is partitioned into two stages. We now describe the first stage.

Stage 1:

1. Whenever possible, apply *remove constant*.
2. Apply *reduce support* until all random variables have support of size at most two. (Different variables may have different support.)
3. Apply *align with 0* to all variables.
4. Apply *merge* until either the number of random variables is reduced to one, or all random variables have mean at least 1/2 (whichever happens first).

Stage 1 must end because with each application of *merge*, the number of random variables decreases. Let $X_1, \ldots, X_{n'}$ be the random variables that we remain with when stage 1 ends. We assume that they are sorted in order of decreasing μ_i. Their number n' may be smaller than n, because some of the transformations remove random variables. These are not arbitrary random variables, as each of them has a support of two values, one of which is 0, and the stopping condition for the merge transformations has been reached. For a random variable X_i as above, let μ_i denote its mean, $(0, v_i)$ its support, and let $s_i = v_i - \mu_i$ denote its *surplus*. Let $s = \sum_{i=1}^{n'} s_i$ denote the total surplus.

PROPOSITION 1. *If the total surplus satisfies $s < \delta$, then $Pr[X \ge \mu+\delta] = 0$.*

PROOF. X is maximized when all X_i come up equal to their respective v_i. In this case

$$X = \sum_{i=1}^{n'} (\mu_i + s_i) = \mu + s < \mu + \delta.$$

□

Hence we may assume without loss of generality that $s \geq \delta$.

LEMMA 2. *If stage 1 ended with a random variable with mean below* $1/2$*, then* $Pr[X < \mu] \geq \delta/(1/2+\delta)$.

PROOF. In this case, exactly one random variable remains. Let X_1 be the random variable left, with mean $\mu_1 < 1/2$ and support $\{0, v_1 = \mu_1 + s\}$. Note that the event $X_1 = 0$ implies $X < \mu$. Now $Pr[X_1 = 0] = s/(\mu_1 + s) \geq \delta/(1/2+\delta)$, because $s \geq \delta$ and $\mu_1 < 1/2$. □

Hence we may also assume that stage 1 ended with all random variables having mean at least $1/2$. The following property will be used in this case.

PROPOSITION 2. *If stage 1 ended with all random variables having mean at least* $1/2$*, then* $\mu_1/2 \leq \mu_{n'} \leq \mu_1 < 3/2$.

PROOF. Recall that the random variables are assumed to be sorted with μ_1 being the largest mean and $\mu_{n'}$ being the smallest mean.

If no random variable has mean greater than 1, then we are done. Hence consider the first time that a random variable with mean greater than 1 is created. This happens by merging two random variables, say X_i and X_j. Let $\mu_i \geq \mu_j$ be their means before the merge. By the definition of *merge*, no other variable had mean smaller than μ_i. By the stopping rule for stage 1, $\mu_j < 1/2$. To get a variable with mean greater than 1, we must have $\mu_i > 1/2$. Note that stage 1 ends after the merge, because no variable with mean below $1/2$ is left. Hence the new variable created becomes X_1 with $1 < \mu_1 < 1+1/2 = 3/2$. But as $\mu_1 \leq 2\mu_i$ and $\mu_{n'} \geq \mu_i$, it follows that $\mu_{n'} \geq \mu_1/2$. □

Let us pause at this point and explain what remains to be proved. All random variables can be assumed to be 2-valued, with one of the values being 0, and with all means μ_i satisfying $\mu_1/2 \leq \mu_i \leq \mu_1$. Moreover, the total surplus s satisfies $s \geq \delta$. For random variables as above we in fact will bound $Pr[X < \mu]$ rather than $Pr[X < \mu+\delta]$. Lemma 3 (its first part) and Lemma 4 will show that $Pr[X < \mu] \geq \min[\delta/(\mu_1+\delta), 1/13]$. This almost proves Theorem 1, except that it might happen that at the end of stage 1, $\mu_1 > 1$. This possibility is handled in the second part of Lemma 3, by showing that one merge operation before the end of stage 1 we had $Pr[X < \mu+\delta] \geq \delta/(1+\delta)$.

The following proposition is used several times in the proofs Lemmas 3 and 4. It is most effective when $s < \mu_n$, and μ_n is not much smaller than μ_1.

PROPOSITION 3. *Let* $X_1, \ldots, X_n$ *be independent random variables with means* $\mu_1 \geq \ldots \geq \mu_n$ *and supports* $\{0, \mu_1 + s_1\}, \ldots, \{0, \mu_n + s_n\}$*, and let* $X = \sum_{i=1}^n X_i$*,* $\mu = \sum_{i=1}^n \mu_i$ *and* $s = \sum_{i=1}^n s_i$*. Then*

$$Pr[X < \mu - \mu_n + s] \geq \frac{s}{\mu_1 + s}$$

PROOF. It suffices that one random variable comes up zero to imply $X < \mu + s - \mu_n$. (The inequality is strict because only a variable with $s_i > 0$ may come up 0.) Hence:

$$Pr[X \geq \mu + s - \mu_n] = \prod_{i=1}^{n} \frac{\mu_i}{\mu_i + s_i} \leq \prod_{i=1}^{n} \frac{\mu_1}{\mu_1 + s_i}.$$

Given that $\sum_{i=1}^n s_i = s$ and that $s_i \geq 0$, the above product is maximized when $s_1 = s$ and $s_i = 0$ for all $i > 1$, giving $\mu_1/(\mu_1 + s)$. Hence $Pr[X < \mu - \mu_n + s] \geq s/(\mu_1 + s)$. □

The following lemma illustrates the desired outcome of stage 1.

LEMMA 3. *If stage 1 ended with all random variables having mean at least* $1/2$*, and if* $s < \mu_{n'}$*, then* $Pr[X < \mu] \geq \delta/(3/2+\delta)$*. If in addition* $Pr[X < \mu] < \delta/(1+\delta)$ *and* $\delta \leq 1/12$ *then one merge operation before the end of stage 1 it must have been the case that* $Pr[X < \mu+\delta] \geq \delta/(1+\delta)$.

Remark: The choice of $\delta \leq 1/12$ in the second part of Lemma 3 is made because $\delta/(1+\delta) = 1/13$ for $\delta = 1/12$. The limiting factor for improving beyond $1/13$ is Lemma 4 rather than Lemma 3. For $\delta > 1/12$ the second part of Lemma 3 simply implies that $Pr[X < \mu+\delta] \geq Pr[X < \mu + 1/12] \geq 1/13$.

PROOF. The surplus s is smaller than the mean of any of the random variables. Using Proposition 3 we then have $Pr[X < \mu] = \frac{s}{\mu_1+s}$. Using the assumption that $s \geq \delta$ and the fact that $\mu_1 \leq 3/2$ (Proposition 2), we have that $Pr[X < \mu] \geq \delta/(3/2+\delta)$.

To prove the second part of the lemma, note that if it happens that $\mu_1 \leq 1$ then we have $Pr[X < \mu] \geq \delta/(1+\delta)$. Hence we may assume that $\mu_1 > 1$, implying in particular that X_1 is the result of the last *merge* operation (see proof of Proposition 2). Let $s' = s - s_1$ be the surplus of all variables except for X_1. Then analysis as above implies that

$$Pr[X \geq \mu] \leq \frac{\mu_2}{\mu_2 + s'} \leq \frac{1}{1+s'}.$$

Hence if $s' \geq \delta$, $Pr[X < \mu] \geq \delta/(1+\delta)$. So we can assume that $s' < \delta$.

Let us backtrack the last merge operation. Hence instead of X_1 we have two variables X_i and X_j that were merged to give X_1. Let their means be $\mu_i \geq \mu_j$, and their surpluses be s_i and s_j. Observe that necessarily $\mu_j < 1/2$ (otherwise the merge operation would not have been performed), and then the assumption that $\mu_1 > 1$ implies that $\mu_i > 1/2$. As the total surplus of all random variables except for X_i and X_j is $s' < \delta$, we must have $X_i + X_j$ come up larger than $\mu_i + \mu_j$ for $X \geq \mu + \delta$. We consider now two cases.

Case 1: $s_i > 2\delta$**.** Then $Pr[X_i = 0] = \frac{s_i}{\mu_i+s_i} \geq \frac{2\delta}{\mu_i+2\delta}$. If $X_i = 0$, then in order to have $X_i + X_j > \mu_i + \mu_j$ we must have $X_j > \mu_i + \mu_j$. But this happens with probability at most $\frac{\mu_j}{\mu_i+\mu_j} \leq \frac{1/2}{\mu_i+1/2}$. Hence

$$Pr[X < \mu+\delta] \geq \frac{2\delta}{\mu_i + 2\delta} \cdot \frac{\mu_i}{\mu_i + 1/2} \geq \frac{\delta}{1+\delta}$$

where the last inequality holds for $\delta \leq 1/2$ because $1/2 \leq \mu_i \leq 1$.

Case 2: $s_i \leq 2\delta$**.** Define $s'' = s' + s_i$ as the surplus of all random variables except for s_j, and observe that $s'' < 3\delta \leq 1/4$, the last inequality holding for $\delta \leq 1/12$.

- If $s'' < \delta$ and $s_j < 1/2$ then it suffices for one random variable to come up 0 to ensure $X < \mu + \delta$. As necessarily $s'' + s_j \geq \delta$ and $\mu_k \leq 1$ for all k, this happens with probability at least $\delta/(1+\delta)$.

- If $s'' < \delta$ and $s_j \geq 1/2$ then it suffices for X_j to come up 0 to ensure $X < \mu + \delta$. This happens with probability at least 1/2.

- If $\delta \leq s'' < \delta + \mu_j$ and $s_j \geq \delta$ then it suffices for X_j to come up 0 to ensure $X < \mu + \delta$. This happens with probability at least $\delta/(1/2 + \delta)$.

- If $\delta \leq s'' < \delta + \mu_j$ and $s_j < \delta$ then it suffices for some random variable other than X_j to come up 0 to ensure $X < \mu + \delta$. (Recall also that $s'' \leq 1/4$.) This happens with probability at least $\delta/(1 + \delta)$.

- If $s'' \geq \delta + \mu_j$ then there is probability of at least $\frac{\delta+\mu_j}{1+\delta+\mu_j}$ for a random variable other than X_j to come up 0. Thereafter X_j must come up at least $\mu_j + \delta + 1/2 - s'' \geq \mu_j + \delta + 1/4$ for $X \geq \mu + \delta$. The probability of this is at most $\mu_j/(\mu_j + 1/4 + \delta)$. Hence

$$Pr[X < \mu + \delta] \geq \frac{\delta + \mu_j}{1 + \delta + \mu_j} \cdot \frac{1/4 + \delta}{\mu_j + 1/4 + \delta} \geq \frac{\delta}{1 + \delta}$$

  where the last inequality holds when $4\delta^2 + 4\mu_j\delta \leq 1$, which is true for our parameters of $\mu_j \leq 1/2$ and $\delta \leq 1/12$.

□

Summarizing, the situation so far is that we may assume that $\mu_1/2 \leq \mu_{n'} \leq \mu_1$ and $s \geq \mu_{n'}$. We shall prove that in this case $Pr[X < \mu] \geq 1/13$. To prove this, we perform stage 2 of our sequence of transformations. It is composed of a modified form of the merge operations, that we call *modified merge*. The modification will allow us to deal with the event $X < \mu$ rather than $X < \mu + \delta$. Recall that the *reduce support* operation was based on a linear program that minimized $Pr[X < \mu+\delta]$ (via the definition of the q_j). Modify the *reduce support* operation by modifying the objective function of the linear program to be $Pr[X < \mu]$ (by making the respective change in the definition of q_j). Use this *modified reduce support* rather than the original *reduce support* as the second step of *modified merge*. Now *modified merge* does not increase $Pr[X < \mu]$.

Note that an application of *modified merge* may result in a random variable whose mean is smaller than $\mu_1/2$. (For simplicity of notation, we assume that after every step the variables are renamed so as to keep μ_1 the largest mean.) However, even with repeated applications of *modified merge*, there will be at most one such random variable. Let us define $s' = \sum s_i$, where the sum is taken over all random variables whose mean is at least $\mu_1/2$. In particular, at the time when stage 1 ends, $s = s'$.

Stage 2. Apply *modified merge* (on the two random variables with currently lowest mean) until either the number of nonconstant random variables is reduced to one, or the condition $s' \leq \alpha\mu_1$ has been reached, for some constant $0 < \alpha < 1/2$ that will be determined later. Stage 2 must eventually end, because with each application of *merge* the number of random variables decreases.

LEMMA 4. *When $\alpha = 1/3$, then either at the time stage 2 ends or one* modified merge *operation before stage 2 ends*

$$Pr[X < \mu] \geq 1/13$$

The proof of Lemma 4 involves a detailed case analysis and appears in Section A in the appendix.

This completes the proof of Theorem 1.

It is straightforward to modify inequality (1) so that there is no formal requirement that the random variables are non-negative, or that their mean is bounded by 1. Let w be the maximum over all random variables $X_1, \ldots, X_n$ of the respective $\mu_i - l_i$, where l_i is the lowest value in the support of X_i. Then

$$Pr[X \leq \mu + \delta w] \geq \min[\delta/(1 + \delta), 1/13] \quad (2)$$

The constant 1/13 in Theorem 1 is not best possible, and can be improved with more detailed case analysis. We suspect that the true constant should be $1/e$. Presumably, the way to prove a tight result is to find a sequence of transformations on the random variables that does not increase $Pr[X < \mu + \delta]$, and that gradually brings them to the conjectured worst case for $[X < \mu + \delta]$. The sequence of transformations performed in our proof of Theorem 1 manages to achieve this only when $\delta \leq 1/12$ (or some other constant not far from 1/12). However, it fails to characterize the worst case for the perhaps more interesting $\delta = 1$. The idea in the proof is to transform the random variables into a situation where a case analysis becomes manageable, at the possible cost of giving up the tightness of the bound. The main principles used are reducing the support of every random variable to two values, getting all random variables (perhaps except one) to have roughly the same mean, reducing the surplus to be of order of magnitude comparable to this mean, and extracting from arbitrarily many random variables a single event of interest, as done in Proposition 3. It should be clear to the reader that more detailed case analysis would provide tighter results. But let us point out some limitations that relate to Lemma 4. As long as one chooses α not larger than $\mu_1/2$ (and in fact, not larger than $3\mu_1/2$), and analyses only the situation at the end of stage 2 or one step earlier, one cannot obtain a bound better than $Pr[X < \mu] \geq 2/9$. For example, assume that during stage 2 we are left with three variables, each with support $(0, \mu)$ and mean $\mu/3$. At this point, $Pr[X < \mu] = (2/3)^3 = 8/27 < 1/e$. After a merge operation, this probability decreases further to $(1/3) \cdot (2/3) = 2/9$. One merge operation later, stage 2 ends. Hence to get (nearly) tight results using the current approach, one may need to modify the definition of stage 2, and perform much more extensive (possibly computer assisted) case analysis.

4. PROOF OF THEOREM 2

The reader is assumed to be familiar with elementary methods in probability (such as the use of Markov's inequality, Chebschev's inequality, Chernoff bounds). If needed, see details in [1], for example.

We query at random t vertices, and obtain their degrees. Let d_i be the degree returned by the ith query. Basically, our estimator for d will be $d^* = \frac{1}{t}\sum_{i=1}^{t} d_i$. In section 4.3 we shall modify this estimator so as to improve its quality. For simplicity of the analysis, we assume that sampling is done with replacement (the same vertex might be queried more than once). This is insignificant when t is small (e.g., $t \leq \sqrt{n}$), though note that for large values of t (and in particular, when $t = n$) sampling without replacement gives better estimates than sampling with replacement.

Note that the expectation of our estimate satisfies

$$E[d^*] = d \tag{3}$$

Hence the estimator is *unbiased*. In deviations from the expectation, we will analyse separately the events $d^* > d$ and $d^* < d$, or rather, $d^* < d/2$.

4.1 The estimate is not too high

Here we shall use Theorem 1. As an immediate consequence of this theorem (taking $\delta = 1$, and using the fact that the degree of a sampled vertex is a nonnegative random variable with expectation d) we have the following corollary.

COROLLARY 1. *There is some universal constant $\alpha > 0$ such that for every graph with average degree d, by querying t random vertices (with replacement) for their degree, the average d^* satisfies $Pr[d^* \leq (1 + 1/t)d] \geq \alpha$.*

We can take $\alpha = 1/13$ in Corollary 1, and we conjecture that the Corollary is also true with $\alpha = 1/e$.

4.2 The estimate is not too low

We assume that the average degree in the graph is at least d_0. Our sampling algorithm queries $t = k\sqrt{n/d_0}$ vertices at random and reports the sum of the degrees. Here k is a parameter that will later be chosen to be of order $1/\epsilon$.

Let X_i be the random variable that denotes the degree of the ith query, and let $X = \sum_{i=1}^{t} X_i$. Then $E[X] = t \cdot d$. We would like to show that the typical value of X is not much smaller than $E[X]/2$. This would follow from Chebyschev's inequality had the variance of X been small compared to $(E[X])^2$. Unfortunately, this is not the case. Vertices of very high degree may cause the variance to exceed $(E[X])^2$. To overcome this problem we observe that in every graph, the vertices of very high degree contain at most slightly more than half the endpoints of the edges. (There can be only few vertices of very high degree, as otherwise the average degree also becomes high. Not having parallel edges, this implies that there are only few edges with both endpoints in vertices of very high degree. All other edges must have at least one endpoint in a vertex whose degree is not very high.) The contribution to X of vertices whose degree is not very high is concentrated around its mean, because for them the variance is small. This explains why the value of X is likely to be above $E[X](1 - \epsilon)/2$. We note however that to get tighter bounds on the number of queries t as a function of ϵ, we take into account also the contribution of vertices of very high degree to X. Based on these principles, in Section B in the Appendix we prove the following Corollary.

COROLLARY 2. *For arbitrary $\lambda > 0$ (that will later be fixed to $50\sqrt{2}/\alpha$, where α is as in Corollary 1), with probability at least $1 - 4\sqrt{2}/\lambda - 2^{-\Omega(\lambda)}$,*

$$X \geq \frac{E[X]}{2}(1 - \frac{\lambda}{k})$$

4.3 Combining the upper and lower bound

Let us set $k = \lambda/\epsilon$, and hence from Corollary 2 we have that with probability at least $1 - 4\sqrt{2}/\lambda - 2^{-\Omega(\lambda)}$,

$$X \geq \frac{E[X]}{2}(1 - \epsilon)$$

By Corollary 1, we have that with probability at least α, $X \leq E[X](1 + 1/t)$. We assume here for simplicity that ϵ is small enough so that $1/(1 - \epsilon) \simeq 1 + \epsilon$. Likewise, t is large enough compared to $1/\epsilon$ so that $(1 + 1/t)/(1 - \epsilon) \simeq 1 + \epsilon$.

An *unbiased estimate* consists of taking t samples and returning their sum X. Perform $l = 2/\alpha$ independent unbiased estimates for X. Our estimation procedure returns $X_{\min}$, the minimum of these estimates. (Equivalently, we set $d^* = X_{\min}/t$.)

$$Pr[X_{\min} \leq E[X](1 + \frac{1}{t})] \geq 1 - (1 - \alpha)^{2/\alpha} \geq 1 - \frac{1}{e^2} \geq \frac{5}{6}$$

$$Pr[X_{\min} \geq \frac{E[X]}{2}(1 - \epsilon)] \geq 1 - \frac{2}{\alpha}(\frac{4\sqrt{2}}{\lambda} + 2^{-\Omega(\lambda)}) \geq 5/6$$

where the last inequality uses $\lambda = 50\sqrt{2}/\alpha$. This gives $k = \lambda/\epsilon < 72/\alpha\epsilon$. The total number of queries used in our estimation procedure is $l \cdot t$. This gives:

COROLLARY 3. *For some universal constant β, using*

$$\beta\frac{\sqrt{n/d_0}}{\epsilon}$$

queries, one can estimate the average degree d of an n node graph within a ratio of $(2 + \epsilon)$, provided that $d > d_0$.

PROOF. Setting $\beta = (\frac{2}{\alpha})(\frac{72}{\alpha}) = 144/\alpha^2$, we perform $2/\alpha$ unbiased estimates, each with $t = 72\sqrt{n/d_0}\alpha^{-1}\epsilon^{-1}$ queries, and take the minimum of the estimations that they give. □

Let us note here the role of Corollary 1. It allows us to substitute a universal constant for α (which is shown to be at least $1/13$ in Theorem 1, though we conjecture that $1/e$ also works). Had we used Markov's inequality instead, we could have taken $\alpha \simeq \epsilon$, loosing a factor of ϵ^{-2} in the number of queries used by the estimation procedure.

The sample size in Corollary 3 is essentially best possible. See Section C in the appendix for details.

5. QUICKLY ESTIMATING THE LOAD ON A NETWORK

We have seen how to estimate the average degree in a graph using a relatively small number of degree-queries. Graph problems are often abstractions of other more concrete problems. As an example (which motivated this study), consider the following problem motivated and studied in [2].

The input is a connected network G with n vertices and m edges (namely, a graph). Between every two vertices there is a shortest path (a path that crosses the smallest number of edges). We assume here that shortest paths are unique, a point that we shall return to later. For an integer parameter c (that may depend on n), we wish to find all edges that are members of at least c shortest paths. In the terminology of [2], these edges are called "weakest links", apparently because these are the edges where failure may cause the largest amount of damage to the performance of the network. Finding all weakest links can be done in time $O(nm)$ using an algorithm for all pairs shortest paths. The goal in [2] is to do better. They propose a randomized algorithm that with high probability, has the following guarantee:

- **Finds weakest links.** It outputs all edges that belong to at least c shortest paths.

- **Avoids false alarms.** It does not output any edge that is a member of less than $(1-\epsilon)c$ shortest paths.

The running time of the algorithm in [2] is $O(\frac{mn^2 \log n}{c\epsilon^2})$, which is better than that of all pairs shortest paths when $c \gg n \log n$. The basic idea in this algorithm is to choose $k \simeq \frac{n^2 \log n}{c\epsilon^2}$ pairs of vertices at random, and for each pair to perform a shortest path computation (taking $O(m)$ operations per-pair). Using the collection of k shortest paths that are found, one estimates in how many shortest paths each edge participates.

Here we present a faster algorithm for finding the weakest links. It is based on two observations. One is that the cost of performing single source shortest path computations (namely, that of finding the shortest paths from one vertex to all other vertices) is $O(m)$, similar to that of finding the shortest path between one pair of vertices. The other observation is that the estimation problem that this gives rise to can be cast as that of estimating the average degree in a graph. The improved running time comes at a cost of a somewhat weaker guarantee in terms of false alarms.

- **Avoids false alarms.** The algorithm does not output any edge that is a member of less than $(1/2-\epsilon)c$ shortest paths.

As in [2] we assume that shortest paths are unique. This requires a convention for breaking ties between paths of equal length. We shall use the same convention that is proposed in [2], namely, to take the lexicographically first such path. See Section D for more details.

PROPOSITION 4. *Under the tie breaking convention specified above, there is an $O(m)$-time algorithm that does the following. Given a connected graph G with n vertices and m edges and an arbitrary vertex v, it simultaneously counts for every edge e, for how many vertices u does edge e participate in the shortest path connecting u and v.*

The proof of Proposition 4 is sketched in Section D in the appendix.

Consider now a particular edge e. It induces the following graph G_e. The vertices of G_e are the vertices of G. Two vertices are connected by an edge in G_e iff e is on their unique shortest path in G. Edge e is on c shortest paths in G iff the average degree in G_e is at least $2c/n$. By Theorem 2, $O(\sqrt{n/d_0}/\epsilon)$ degree queries suffice in order to estimate the average degree in a graph with average degree at least d_0. To make the probability of error in this estimation below $1/n^2$, one can repeat the estimation procedure $O(\log n)$ times, and take the median of the estimations. We shall set $d_0 = (1-\epsilon)c/n$. Now observe that Proposition 4 implies that we can simultaneously obtain the degree of v in the graphs G_e for all e, and all this in time $O(m)$. Hence using $k = O(\frac{\log n \sqrt{n/(c/n)}}{\epsilon}) = O(\frac{n \log n}{\epsilon\sqrt{c}})$ single source shortest path computations one can with high probability find all weakest links (edges that are on more than c shortest paths), and avoid any false alarms (by edges that are on less than $(1/2-\epsilon)c$ shortest paths). This proves Theorem 3.

Theorem 3 offers a saving of roughly $n/\sqrt{c}$ in the running time compared to the running time of $O(\frac{mn^2 \log n}{c\epsilon^2})$ in [2]. (Note however that ϵ has different meanings in the two bounds. Hence the saving comes at the cost of allowing more false alarms.)

Acknowledgements

I thank Johan Hastad, Michael Langberg, Eran Ofek, Gideon Schechtman, Benjy Weiss and Avi Wigderson for their interest and involvement in various stages of this work.

6. REFERENCES

[1] N. Alon and J. Spencer. *The Probabilistic Method.* Wiley-Interscience, 2000.

[2] N. Devanur, R. Lipton, N. Vishnoi. "Who's the weakest link?". *Second Symposium on Stochastic Algorithms, Foundations and Applications*, SAGA 2003.

[3] L. Dubins and L. Savage. *Inequalities for Stochastic Processes. (How to gamble if you must.)* Dover Publications, New York, 1976.

[4] O Goldreich and D. Ron. "On estimating the average degree of a graph". *Electronic Colloquim on Computational Complexity (ECCC)*, TR04-13, 2004.

[5] A. Siegel. "Median bounds and their application". *Journal of Algorithms*, 38(1), 184–236, 2001.

APPENDIX

A. ANALYSIS OF STAGE 2

Here we prove Lemma 4.

LEMMA 5. *If stage 2 ends without the condition $s' \leq \alpha\mu_1$ being reached, then $Pr[X < \mu] > \alpha/(1+\alpha)$.*

PROOF. In this case we have only one nonconstant random variable, X_1, with support $\{0, \mu_1 + s'\}$.

$$Pr[X < \mu] = Pr[X_1 = 0] = \frac{s'}{\mu_1 + s'} > \frac{\alpha}{1+\alpha}.$$

□

LEMMA 6. *If stage 2 ends with $\beta\mu_1 \leq s' \leq \alpha\mu_1$, where $0 \leq \beta \leq \alpha$ is some constant that will be optimized later, then*

$$Pr[X < \mu] \geq \min\Big[\Big(\frac{\alpha - 2\alpha^2}{1+\alpha}\Big),\ \Big(\frac{\beta - 2\beta^2}{1+\beta}\Big)\Big].$$

PROOF. Consider first only the random variables with mean at least $\mu_1/2$, let X' be their sum and let μ' be the expectation of X'. Over these random variables, the surplus is $s' = \gamma\mu_1$, with $\beta \leq \gamma \leq \alpha < 1/2$. By Proposition 3,

$$Pr[X' < \mu' - (1/2-\gamma)\mu_1] \geq \frac{s}{\mu_1 + s} = \frac{\gamma}{1+\gamma}.$$

The event $X' < \mu' - (1/2-\gamma)\mu_1$ does not yet imply that $X < \mu$. There still might be one variable $X_{n'}$ with $\mu_{n'} < \mu_1/2$. If $X_{n'}$ turns out $\mu_{n'} + s_{n'}$ and $s_{n'} \geq (1/2-\gamma)\mu_1$ then it still may hold that $X \geq \mu$.

Let us first assume that $\mu_{n'} \leq s' = \gamma\mu_1$. Then by Markov's inequality,

$$Pr[X_{n'} \geq \mu_{n'} + s_{n'}] \leq \frac{\mu_{n'}}{\mu_{n'} + s_{n'}} \leq \frac{\gamma\mu_1}{\gamma\mu_1 + (1/2-\gamma)\mu_1} \leq 2\gamma.$$

Hence

$$Pr[X < \mu] \geq \frac{\gamma}{1+\gamma} \cdot (1 - 2\gamma) = \frac{\gamma - 2\gamma^2}{1+\gamma}.$$

For $0 < \beta \leq \gamma \leq \alpha < 1/2$, the expression above is minimized when $\gamma \in \{\alpha, \beta\}$.

We are left with the case that $\mu_{n'} > s'$. But then we have

$$Pr[X < \mu] \geq Pr[X_{n'} = 0] \geq \frac{s_{n'}}{\mu_{n'} + s_{n'}} \geq \frac{1/2 - \gamma}{1 - \gamma}$$

where we have used the facts that $\mu_{n'} < \mu_1/2$ and $s_{n'} \geq (1/2 - \gamma)\mu_1$. As $\gamma \leq \alpha$, we have that $Pr[X < \mu] > (1/2 - \alpha)/(1-\alpha)$. But this probability is larger than $(\alpha - 2\alpha^2)/(1+\alpha)$ of the previous case, and hence can be ignored. □

LEMMA 7. *If stage 2 ends with $s' < \beta\mu_1$, and $0 < \beta < \alpha/2$, then one merge prior to the end of stage 2 it must have been the case that $Pr[X < \mu]$ was at least the minimum of the following expressions:*

1. $\frac{\alpha-\beta}{1/2+\alpha-\beta} \cdot \frac{1/2-\beta}{1-\beta}$
2. $\frac{\alpha-3\beta/2}{1+\alpha-3\beta/2} \cdot \frac{1-3\beta/2}{3/2-3\beta/2}$
3. $\frac{\alpha-2\beta}{1+\alpha-2\beta}$
4. $\left(\frac{1/2-\beta}{1-\beta}\right)^2$
5. $\frac{1/2-3\beta/2}{3/2-\beta} \cdot \frac{1-3\beta/2}{3/2-\beta}$

PROOF. Consider the last two random variables to have been merged, say X_i and X_j, with means $\mu_i \geq \mu_j$, and let μ'_1 be the largest mean at the time. After the *modified merge* of X_i and X_j, the largest mean μ_1 may still have been μ'_1, but it could also be as high as $\mu_i + \mu_j$, if this happens to be higher than μ'_1. In fact, μ_1 may also be lower than μ'_1, if only one variable is left at the end of stage 2, and this variable underwent an *align with 0* operation. However, in this case the bounds that we get for $X < \mu$ are much stronger than what we get otherwise (details omitted), so we shall ignore this case.

We analyse the situation one merge operation before the end of stage 2. Note that we know that at that time, $s' \geq \alpha\mu_1$, because otherwise stage 2 would have ended earlier. Likewise, the sum $\sum s_i$ taken over all variables except X_i and X_j is at most $\max[\beta\mu'_1, \beta(\mu_i + \mu_j)]$, because otherwise we could not have had $s' < \beta\mu_1$ at the end of stage 2. We consider now two cases.

Case 1: $\mu_j < \mu'_1/2$. Hence μ_j did not contribute to s'. Note that $\mu_j \geq mu'_1/2$ and follows that $s_i \geq \alpha\mu'_1 - \beta\mu_1$. Hence at that point,

$$Pr[X_i = 0] \geq \frac{\alpha\mu'_1 - \beta\mu_1}{\mu_i + \alpha\mu'_1 - \beta\mu_1}$$

If $X_i = 0$ then in order to have $X \geq \mu$, X_j must contribute at least $\mu_i - \beta\mu_1$ beyond μ_j to X. This may happen with probability at most $\mu_j/(\mu_j + \mu_i - \beta\mu_1)$. We then have

$$Pr[X < \mu] \geq \frac{\alpha\mu'_1 - \beta\mu_1}{\mu_i + \alpha\mu'_1 - \beta\mu_1} \cdot \frac{\mu_i - \beta\mu_1}{\mu_j + \mu_i - \beta\mu_1}$$

The above expression is minimized when μ_j is maximized (note that increasing μ_j may allow us to increase μ_1, though we are not forced to do so), namely, when $\mu_j = \mu'_1/2$. As $\mu_j \geq \mu'_1/2$, it follows that $\mu_i + \mu_j \geq \mu'_1$. The expression above is minimized when μ_1 is maximized, namely, $\mu_1 = \mu_j + \mu_i]$. Normalising μ'_1 to 1, and keeping the notation μ_i to denote μ_i/μ'_1, we have after some rearrangements

$$Pr[X < \mu] \geq \frac{\alpha - \beta/2 - \beta\mu_i}{(1-\beta)\mu_i + \alpha - \beta/2} \cdot \frac{(1-\beta)\mu_i - \beta/2}{(1-\beta)\mu_i + 1/2 - \beta/2}$$

The expression above is defined for all $\mu_i \geq 0$. It equals 0 for $\mu_i = \{(\alpha - \beta/2)/\beta, \beta/2(1-\beta)\}$ and positive in between. Moreover, there are only two points where the derivative with respect to μ_i of this expression vanishes (as it is a ratio of two quadratics), and for $\beta < 2\alpha/3$ the expression is positive in the allowed range of $1/2 \leq \mu \leq 1$. It follows that the expression is minimized when $\mu_i \in \{0, 1, \}$, giving

$$Pr[X < \mu] \geq$$

$$\min\left[\frac{\alpha-\beta}{1/2+\alpha-\beta} \cdot \frac{1/2-\beta}{1-\beta}\ ,\ \frac{\alpha-3\beta/2}{1+\alpha-3\beta/2} \cdot \frac{1-3\beta/2}{3/2-3\beta/2}\right]$$

This gives items 1 and 2 of the lemma.

Case 2: $\mu_j \geq \mu'_1/2$. Hence both s_i and s_j did contribute to s' (before the last merge), and moreover, $\mu_i + \mu_j \geq \mu'_1$. As in case 1, the worst possibility here is that $\mu_1 = \mu_i + \mu_j$. To simplify notation and without loss of generality we may assume that $\mu'_1 = 1$, and then $1/2 \leq \mu_j \leq \mu_i \leq 1$. We have that $s_i + s_j > \alpha - \beta(\mu_i + \mu_j) \geq \alpha - 2\beta$. We are guaranteed than $X < \mu$ if $X_i + X_j < \mu_i + \mu_j - \beta(\mu_i + \mu_j)$. Hence let B denote the event $[X_i + X_j < (1-\beta)(\mu_i + \mu_j)]$. We perform now a subcase analysis for $Pr[B]$.

1. It suffices that either $X_i = 0$ or $X_j = 0$ for B to hold. In this case, using $\mu_i \leq \mu_1$, $s_i + s_j \geq \alpha - 2\beta$, Proposition 3 implies that

   $$Pr[B] \geq \frac{\alpha - 2\beta}{1 + \alpha - 2\beta}$$

   This gives item 3 in the statement of the lemma.

2. B holds iff $X_i = \mu_i + s_i$. In this subcase necessarily $\mu_i + s_i \geq (1-\beta)(\mu_i + \mu_j)]$. Using the fact that $\mu_i \leq 1$ and $\mu_j \geq 1/2$ we have

   $$Pr[B] \geq \frac{1/2 - 3\beta/2}{3/2(1-\beta)}$$

   This subcase is dominated by the subcase above and hence can be ignored.

3. B holds iff $X_j = \mu_j + s_j$. This subcase is dominated by the subcase above and can be ignored.

4. B holds unless both $X_i = 0$ and $X_j = 0$. Then necessarily $\mu_i + s_i \geq (1-\beta)(\mu_i + \mu_j)$ and $\mu_j + s_j \geq (1-\beta)(\mu_i + \mu_j)$. We have

   $$Pr[B] \geq$$

   $$\left(\frac{(1-\beta)(\mu_i + \mu_j) - \mu_i}{(1-\beta)(\mu_i + \mu_j)}\right) \cdot \left(\frac{(1-\beta)(\mu_i + \mu_j) - \mu_j}{(1-\beta)(\mu_i + \mu_j)}\right)$$

   For fixed $\mu_i + \mu_j$ this expression is minimized when $\mu_i - \mu_j$ is maximized. Hence either $\mu_i = 1$ or $\mu_j = 1/2$. Thereafter, it can be verified that the expression is minimized when the other mean is either maximized or minimized, giving us three possible local minimum points, $\mu_i, \mu_j \in \{1/2, 1\}$, $\mu_j \leq \mu_i$. Two of these give identical values (the cases that $\mu_i = \mu_j$), hence we obtain

   $$Pr[B] \geq \min\left[\left(\frac{1/2-\beta}{1-\beta}\right)^2\ ,\ \left(\frac{1/2-3\beta/2}{3/2-\beta} \cdot \frac{1-3\beta/2}{3/2-\beta}\right)\right]$$

   This gives items 4 and 5 in the statement of the lemma.

□

Summing up, we see that after stage 2, $Pr[X < \mu]$ is at least the smallest of the following quantities (where $0 < \beta < \alpha/2 < 1/4$):

- $\frac{\alpha-2\alpha^2}{1+\alpha}$
- $\frac{\beta-2\beta^2}{1+\beta}$
- $\frac{\alpha-\beta}{1/2+\alpha-\beta} \cdot \frac{1/2-\beta}{1-\beta}$
- $\frac{\alpha-3\beta/2}{1+\alpha-3\beta/2} \cdot \frac{1-3\beta/2}{3/2-3\beta/2}$
- $\frac{\alpha-2\beta}{1+\alpha-2\beta}$
- $\left(\frac{1/2-\beta}{1-\beta}\right)^2$
- $\frac{1/2-3\beta/2}{3/2-\beta} \cdot \frac{1-3\beta/2}{3/2-\beta}$

Choosing (suboptimally) $\alpha = 1/3$ and $\beta = 1/8$ gives $Pr[X < \mu] \geq 1/13$ in all cases.

This completes the proof of Lemma 4.

B. BOUNDING THE ESTIMATE ON THE DEGREE FROM BELOW

Partition the set of vertices of G into two sets, H (for high) and L (for low). For a constant c (independent of n, d, k) that will be determined later, the set H contains the $c\sqrt{nd}/k$ vertices of highest degree (breaking ties arbitrarily). The set L contains the other vertices. Every edge has two endpoints. Let us partition the endpoints of edges into the following four sets:

- $E_{H,L}$ (the endpoints in H of edges between H and L)
- $E_{L,H}$ (the endpoints in L of edges between H and L)
- $E_{H,H}$ (the endpoints in H of edges between H and H)
- $E_{L,L}$ (the endpoints in L of edges between L and L)

Observe that $|E_{H,H}| \leq |H|^2 = c^2nd/k^2$. It will be the case that c is a universal constant whereas $k \geq \Omega(1/\epsilon)$, and hence $|E_{H,H}| = O(\epsilon^2 nd)$. Moreover, we allow an error of $\epsilon \cdot nd$ in our estimation of nd. Hence, $E_{H,H}$ has only a low order effect on the accuracy of the estimation. So as to simplify notation and the presentation, we shall simply assume that $|E_{H,H}| = 0$. We shall not give a rigorous proof that this assumption has only a low order effect on our analysis, but merely note here that formalists may redo the analysis without assuming that $|E_{H,H}| = 0$, and at worst this will effect some constants the are eventually hidden by the O notation.

Let $m_1 = |E_{H,L}|$, $m_2 = |E_{L,H}|$ and $m_3 = |E_{L,L}|$. Hence $m_1 + m_2 + m_3 = dn$. Note that $m_2 = m_1$, because $|E_{L,H}| = |E_{H,L}|$. Let us break the random variable X into the sum of three random variables $X = Y_1 + Y_2 + Y_3$, according to the contribution to X from m_1, m_2 and m_3 respectively. Let h denote the minimum degree of a vertex in H.

PROPOSITION 5. *With probability* $1 - 2^{-\Omega(c)}$,

$$Y_1 \geq ch/2$$

PROOF. The expected number of vertices queried from H is $t|H|/n = k\sqrt{n/d_0} \cdot c\sqrt{d_0 n}/kn = c$. With probability $1 - 2^{-\Omega(c)}$, the actual number of vertices queried from H is at least $c/2$. Each such vertex contributes at least h to Y_1. □

PROPOSITION 6. *A vertex in L can cover at most $|H| = c\sqrt{nd_0}/k$ endpoints in $E_{L,H}$.*

PROOF. For every endpoint in $E_{L,H}$ covered by a vertex in L, the other endpoint of the respective edge is in H. As the original graph is a simple graph with no parallel edges, the proof follows. □

PROPOSITION 7. *For $\lambda > 0$, with probability at least $1 - 1/\lambda^2$,*

$$Y_2 \geq E[Y_2] - \lambda\sqrt{cdn/2}$$

PROOF. The variance of Y_2 is maximized if the endpoints of $E_{L,H}$ are concentrated on $m_2/|H|$ vertices (each covering $|H|$ endpoints). Hence:

$$var[Y_2] \leq |H|^2 \frac{m_2}{n|H|} t = cm_2 \leq cdn/2$$

The proof now follows from Chebyschev's inequality. □

PROPOSITION 8. *For $\lambda > 0$, with probability at least $1 - 1/\lambda^2$,*

$$Y_3 \geq E[Y_3] - \lambda\sqrt{\frac{hkm_3}{\sqrt{d_0 n}}}$$

PROOF. The maximum degree of any vertex in L is h. Hence the graph induced by the edges $E_{L,L}$ also has maximum degree at most h. Thus

$$var[Y_3] \leq h^2 \frac{m_3}{h} \frac{t}{n} = hm_3 k/\sqrt{d_0 n}$$

The proof now follows from Chebyschev's inequality. □

PROPOSITION 9. *With probability at least $1 - 2/\lambda^2 - 2^{-\Omega(c)}$,*

$$X \geq \frac{E[X]}{2} + \frac{ch}{2} - \lambda\sqrt{\frac{cdn}{2}} - \lambda\sqrt{\frac{hkm_3}{\sqrt{d_0 n}}} + \frac{km_3}{2\sqrt{d_0 n}}$$

PROOF. $X = Y_1 + Y_2 + Y_3$. By propositions 5,7 and 8 we have that with probability at least $1 - 2/\lambda^2 - 2^{-\Omega(c)}$,

$$X \geq E[Y_2] + E[Y_3] + \frac{ch}{2} - \lambda\sqrt{\frac{cdn}{2}} - \lambda\sqrt{\frac{hkm_3}{\sqrt{d_0 n}}}$$

As $E[Y_1] = E[Y_2]$, we have that $E[X]/2 = E[Y_2] + E[Y_3]/2$. Using $E[Y_3] = m_3 t/n = km_3/\sqrt{d_0 n}$ the proof follows. □

Fix $c = 4\lambda^2$. Then

$$\frac{ch}{2} \cdot \frac{km_3}{2\sqrt{d_0 n}} \geq \left(\lambda\sqrt{\frac{hkm_3}{\sqrt{d_0 n}}}\right)^2$$

implying

$$\frac{ch}{2} - \lambda\sqrt{\frac{hkm_3}{\sqrt{d_0 n}}} + \frac{km_3}{2\sqrt{d_0 n}} \geq 0.$$

The term $\lambda\sqrt{\frac{cdn}{2}} = \lambda^2\sqrt{2dn}$ is at most $\frac{E[X]}{2} 2\lambda^2\sqrt{2}/k$, because $E[X] = dt = k\sqrt{nd}\sqrt{d/d_0}$. Renaming $2\sqrt{2}\lambda^2$ by λ, we obtain Corollary 2.

C. OPTIMALITY OF SAMPLE SIZE

PROPOSITION 10. *For every (reasonable) n, d, ϵ, one can construct a graph G_1 with $(1+\epsilon)nd$ edges and a graph G_2 with $dn/2$ edges, such that $\Omega(\sqrt{n/d}\epsilon^{-1})$ vertices need to be queried in order to have probability above $2/3$ of distinguishing between them.*

PROOF. Graph G_1 has a set A of $\epsilon\sqrt{nd}$ vertices of degree $(1+\epsilon)\sqrt{nd}/\epsilon$, and a set B of $(1+\epsilon)\sqrt{nd}/\epsilon$ vertices of degree $\epsilon\sqrt{nd}$ (e.g., arranged as a complete bipartite subgraph between A and B). The other vertices have degree 0. Graph G_2 has a set C of $\sqrt{nd}/\epsilon$ vertices of degree $\epsilon\sqrt{nd}$.

We sketch the proof of why $\Omega(\sqrt{\frac{n}{d}}\epsilon^{-1})$ queries are necessary. Assume that the number of queries is $\sqrt{\frac{n}{d}}\epsilon^{-1}$. Then there is constant probability that no vertex from A is queried, and the expected number of vertices queried from B is $\epsilon^{-2} + \epsilon^{-1}$. The expected vertices queried from C is ϵ^{-2}. As the standard deviation is of order $\sqrt{\epsilon^{-2}} = \epsilon^{-1}$, there is constant probability that G_1 and G_2 will be confused. □

The optimality of the sample size was proved under the assumption that the only information used by the estimation algorithm is the degree of the queried vertices. More generally, one may think of randomized estimation algorithms that make use of additional information. For example, when querying a vertex of positive degree, the next vertex to query may be chosen at random from the list of neighbors of the current vertex. The use of a more general class of random estimation algorithms may allow either quicker or more accurate estimation of the average degree in a graph. See [4], for example. However, let us explain here some the the advantages of "degree only" sampling, advantages that might be lost by other estimation algorithms.

1. All queries can be made in parallel, which in some contexts results in saving time.
2. Sampling can be done anonymously. The estimation algorithm need not know the identity of queried vertices, nor the identity of their neighbors. Privacy issues may sometimes require that this be the case. (For example, vertices of a graph may represent persons in some community, and an edge may represent some sort of interaction that took place between the respective persons. Persons may be willing to fill an anonymous questionnaire stating with how many different persons they had interaction (namely, their degree), but may not be willing to disclose with whom they had interaction.)
3. In Section 5 there are several different graphs G_e defined on the same set of vertices, and in a single degree query one gets the degrees of the respective vertex in all graphs simultaneously. In order to efficiently estimate the average degree in all graphs, it is useful to have an estimation algorithm for which the choice of which vertex to query does not depend on the graph in question.

D. FINDING SHORTEST PATHS

We first explain the tie breaking convention in more details. We assume that vertices are numbered from 1 to n. A path can be viewed as a sequence of vertices in a natural way. Hence a path is a sequence of numbers. In fact, two sequences correspond to the same path, depending on which of its two endpoints is considered to be the head of the path, and which is considered to be the tail. The name of the path is taken to be the lexicographically smaller of the two. Given two different paths that connect the same pair of vertices, if they are of equal length we use the convention that the one with the lexicographically smaller name is considered to be shorter.

Now we sketch the proof of Proposition 4.

We assume a model of computation in which algorithms such as single source shortest path take $O(m)$ time. In particular, some basic operations (such as comparison between two $O(\log n)$-bit words) take unit time.

Given a starting vertex v, the distances to all other vertices in G can be computed in $O(m)$ time using breadth first search (BFS). The BFS tree rooted at v gives also shortest paths from v to all vertices. It is quite straightforward to also count for each edge in the BFS tree (starting from edges furthest from the root and moving towards the root) in how many shortest paths (starting from v) it participates. The counting requires only $O(n)$ operations, as there are only $n-1$ edges in the BFS tree.

In general, several different BFS trees can be constructed starting at the same vertex v, depending on the order in which vertices are considered. We shall need to construct two such trees. The *forward tree* rooted at v (gives the lexicographically first shortest paths when v is the first vertex of the path) is constructed using the following rules: the neighbors of every vertex are always scanned in lexicographic order, and when trying to identify and connect to the vertices in the next level of the BFS tree, the vertices at the current level are scanned in the order under which they were first discovered. The *backward tree* rooted at v (gives the lexicographically first shortest paths when v is the last vertex of the path) is constructed using the following rule: for every vertex discovered at level i keep a pointer to the lexicographically first vertex of level $i-1$ that connects to it. Both the forward tree and the backward tree can be constructed in $O(m)$ time.

Given both the forward tree and the backward tree for a vertex v, and using the convention that for vertices lexicographically smaller than v one uses the backward tree and for vertices lexicographically higher than v one uses the forward tree, one can simultaneously count in $O(n)$ time how many shortest paths with an endpoint at v pass through every edge. (Note that this count is 0 for all but at most $2n-2$ edges of the two BFS trees.)