Nonlinear Manuscript No.:
(will be inserted by the editor)

L. Massoulié · E. Le Merrer · A.-M. Kermarrec · A.J. Ganesh

Peer counting and sampling in overlay networks: random walk methods

Abstract We address the problem of counting the number of peers in a peer-to-peer system, and more generally of aggregating statistics of individual peers over the whole system. This functionality is useful in the design of several applications. It is delicate to achieve when nodes are organised in an overlay network, and each node has only a limited, local knowledge of the whole system.

We propose two generic techniques to solve this problem. First, the *Random Tour* method is based on a continuous time random walk, and exploits the return time of the walk to the node originating the query.

Second, the *Sample and Collide* method essentially relies on a sampling sub-routine which returns randomly chosen peers. Such a sampling sub-routine is of independent interest. It can be used for instance for neighbour selection by new nodes joining the system. We use a continuous time random walk to obtain such samples.

The core of the second method consists in gathering samples until a target number of redundant samples are obtained. This method is inspired by the "birthday paradox" technique of [7], upon which it improves by achieving a target variance with fewer samples.

We analyse the complexity and accuracy of the two methods. We illustrate in particular how *expansion* properties of the overlay affect their performance.

We use simulations to evaluate their performance in dynamic environments with sudden changes in peer populations. Both methods track varying system sizes accurately. As predicted by the analysis, the cost incurred by the Sample and Collide method is significantly lower for a comparable accuracy.

L. Massoulié
Microsoft Research, Cambridge, UK
E-mail: lmassoul@microsoft.com

E. Le Merrer
France Telecom R&D, Lannion, France
E-mail: erwan.lemerrer@orange-ft.com

A.-M. Kermarrec
INRIA/IRISA, Rennes, France
E-mail: Anne-Marie.Kermarrec@irisa.fr

A.J. Ganesh
Microsoft Research, Cambridge, UK
E-mail: ajg@microsoft.com

1 Introduction

Peer-to-peer systems have achieved fast and widespread adoption for both legal and illegal applications, ranging from file sharing (e.g. Kazaa or eDonkey) to VoIP (e.g. Skype). It is reasonable to expect novel applications to appear, and the scale of such systems to increase beyond millions of interacting peers.

A key feature of peer-to-peer systems is their *distributed* nature. Indeed, popular architectures organise the peers in an *overlay* network, typically layered over the Internet, and let peers communicate solely with their overlay neighbours. In such architectures, a peer's knowl-

edge of the system is limited to its collection of neighbours. These architectures have good scalability properties; in particular they do not suffer from central servers becoming performance bottlenecks, or single points of failure.

On the flip side however, overlay architectures make it delicate to monitor system characteristics of interest that would be straightforward to observe in centralised systems. One example of such a system characteristic that will concern us in this paper is the system size, namely the number of peers. More generally, we are interested in (approximately) counting the number of peers with given characteristics, or aggregating characteristics of interest of individual peers over all peers.

The need to perform such peer counting arises in the following contexts. Recently proposed overlay maintenance protocols, such as Viceroy [28], rely on approximate knowledge of the overlay size to incorporate a newly arrived peer in the system. Several gossip-based information dissemination protocols (see e.g. [16], [14]) rely on system size to determine the number of gossip targets per peer. We expect that such peer counting could find other applications. For instance, in a Live Media streaming system such as that of [36], it may be of interest to measure the number of peers using a Broadband connection or a dialup connection, in order to decide whether new dialup users can be accepted without compromising performance.

For particular overlay architectures, specific overlay properties may be exploited to efficiently measure system size. In contrast, our aim here is to design measurement techniques that are generic, in that they are applicable to arbitrary overlay networks. We propose two such techniques in this paper.

The first technique, which we call the *Random Tour* method, relies on launching a message from the peer initiating the measurement. This message then performs a random walk along the overlay until it returns to the initiator, i.e. a random tour. Local statistics accumulated along the tour within this message provide an estimate of system size. We analyse the quality of the resulting estimate, and in particular show how the *spectral gap* of the overlay graph, which is in turn affected by the *expansion* properties of the graph, conditions the accuracy of the estimate. This technique can be easily adapted to estimate aggregate statistics of other node properties.

We then propose a second technique, which we call the *Sample&Collide* method. One building block of the method is a sampling function, which aims to provide a requesting overlay node with another node chosen uniformly at random from the overlay. Previous proposals have relied on a Discrete Time Random Walk stopped after a large constant time; clearly, this yields samples biased towards high-degree nodes. We describe a sampling algorithm based on a Continuous Time Random Walk (CTRW), which yields unbiased samples. We characterise the sampling quality/complexity trade-off and find that it is again critically affected by the expansion properties of the overlay graph.

The Sample&Collide method produces an estimate of the system size based on the number of uniform random samples it takes before a target number of redundant samples are obtained. This method is inspired by the "Inverted Birthday Paradox" method of [7]. We provide a detailed analysis of the accuracy and complexity of the Sample&Collide method, and show how it improves upon the original proposal of [7] by achieving a target accuracy with fewer samples.

While we do not study the impact of churn analytically, we evaluate it through simulations. These show that the proposed techniques are robust to both gradual and sudden changes in system size.

The paper is structured as follows: in the next section we survey related work. We present the Random Tour method and its analysis in Section 3, and the Sample&Collide method and its analysis in Section 4. The evaluation of the two methods by simulation is shown in Section 5. We conclude in Section 6.

2 Related Work

We distinguish two classes of techniques for system size estimation. Techniques of the first type are tailored to a specific overlay architecture, while those of the second type are generic and applicable to any overlay.

2.1 Architecture-specific techniques

In structured peer to peer overlay networks, peers are assigned identifiers drawn uniformly at random. The approach taken in [11] deduces the network size in distributed hash tables (DHT) by measuring the density of identifiers around a node initiating a size estimate. The communication cost for getting a relative error of order ϵ is $O(1/\epsilon^2)$ message exchanges, irrespective of the number of nodes N. A similar approach is also considered in [33, 24, 19, 29].

In [13], an estimate of the system size is constructed based on observations of node degrees, and relies on prior knowledge of a power law structure for the distribution of node degrees. A conceptually similar approach is described in [7]. It produces an estimate of system size based on node degree observations, assuming a specific topology (namely, the Erdős-Rényi random graph model). No error estimates are provided in these papers; the cost of the latter is $O(\log N)$.

Another approach involves building a spanning tree on top of the overlay, and using it to estimate the system size [9, 32, 25] by aggregating estimates along the tree. The obtained estimates are then exact in the absence of failures, and the cost is $\Theta(N)$.

Jelasity and Preuß [21] estimate the network size by observing the renewal of contacts in peers' views, in a gossip based overlay.

2.2 Generic techniques

Jelasity and Montresor [20] have considered the following gossip-based method. Initially one distinguished node sets a counter to 1 while all other nodes set their counter to 0. Nodes communicate asynchronously; when a pair of nodes communicates, they both reset their individual counters to the mean of the two previous values. In the long run, all counter values coincide with the reciprocal of the system size. This approach is suitable in stable environments. As all users eventually share the same size estimate, its cost is amortized over all nodes when they are all interested in obtaining such an estimate. A theoretical evaluation of the cost of such schemes can be found in [10]. The cost, evaluated in number of messages, is $\tilde{O}(N^{1+2/d})$ for d-dimensional random geometric graphs.[1] It is $O(N \log(N))$ for expander graphs.

Another generic approach [15, 33, 24], sometimes referred to as probabilistic polling, consists of a querying node requesting all nodes to report their presence probabilistically, the probability of responding being a function of node characteristics, such as distance (in number of hops) from the initial requestor. This produces unbiased estimates. One drawback of the method is that the initial querior is potentially faced with "ACK implosion". The cost of this method scales linearly with system size.

Finally, Bawa et al. [7] propose a method which assumes one can sample peers uniformly at random. They form an estimate of system size based on the number of samples required before the same peer is sampled twice. The cost, measured in number of samples, scales like $\ell\sqrt{N}$ where N is the system size, and for a target relative error of $1/\sqrt{\ell}$. The second technique we shall describe builds on this work. It improves upon it by proposing a scheme to generate approximately uniform random samples, and also reduces the number of samples required to achieve the same target accuracy to $\sqrt{\ell N}$, hence a reduction by a factor of $\sqrt{\ell}$.

[1] See [10] for a definition of such graphs. Here, we write $f(N) = \tilde{O}(g(N)$ when $f(N) = O(g(N)\log(N)^{\beta})$ for some β.

3 The Random Tour method

Here and in the sequel, we assume that peers (or nodes) are organised in an undirected graph, each node being aware of, and able to communicate directly with its neighbours. The node set is denoted by $\mathcal{N}$, and its size is N. The *degree* of node i is denoted d_i, and is by definition the number of neighbours of node i in the overlay graph.

The aim is to design lightweight techniques for estimating for instance the system size N. In fact, our techniques also apply to the estimation of sums of functions of the nodes, $\Phi := \sum_{j\in\mathcal{N}} \phi(j)$, for general functions ϕ. Estimation of the system size is just one special case, corresponding to $\phi \equiv 1$. One may for instance be interested in evaluating the number of nodes j with a degree larger than 100; this would correspond to the special case $\phi(i) = 1$ if $d_i > 100$, and $\phi(i) = 0$ otherwise. There are many variants of interest, e.g. counting peers that have an upload capacity above 10Mb/s. We assume throughout that the graph is connected; if it is not, each node will only be able to estimate the size of its connected component.

This section is organised as follows. We first decribe the Random Tour method. We then prove that it is unbiased. We next interpret it in terms of Continuous Time Random Walks. This allows us to evaluate its accuracy, captured by the variance, and how it is affected by the *spectral gap* of the graph. Finally we discuss the role of the expansion property of the overlay, and the cost/accuracy trade-off of the method.

3.1 Basic Algorithm

An estimation procedure is launched at an arbitrary node, say node i, of the system. It proceeds as follows:

1. The initiator, node i, initialises a counter value, X, to $\phi(i)/d_i$. It forwards a message, tagged with the counter value X, and its identity, i, to one of its neighbours, chosen uniformly at random.
2. A node j, when receiving such a message, if it is not the originating node ($i \neq j$), increments the counter by $\phi(j)/d_j$ ($X \leftarrow X + \phi(j)/d_j$), and forwards it to one of its neighbours, chosen uniformly at random.
3. When the originator, i, receives the message it originally sent, with associated counter value X, it forms the following estimate $\widehat{\Phi}$ of the system size Φ:

$$\widehat{\Phi} = d_i X.$$

The estimate is thus obtained by adding a specific amount to a probe message at each node along a random tour, that is, a random walk started at initiator node i and ended upon return to node i.

3.2 Lack of bias

We now establish that the above estimation procedure is unbiased. We shall use $\mathbf{E}_i(\cdot)$ to denote mathematical expectation when the random walk is started at node i. Note that so far we are considering discrete time random walks.

Proposition 1 *The Random Tour algorithm produces an unbiased estimate, that is, the expectation* $\mathbf{E}_i(\Phi)$ *coincides with the quantity to be estimated,* $\sum_{j\in\mathcal{N}} \phi(j)$.

Proof: Consider the discrete time random walk started at node i, which, from a given state j goes to a randomly chosen neighbour of j, each neighbour being equally likely (and thus chosen with a probability of $1/d_j$). Denote by Y_n the position of the random walk at time step n. Then (Y_n) is a Markov chain with transition probabilities $p_{jk} = 1/d_j$ if k is a neighbour of j and 0 otherwise. If the graph is connected, then the transition probability matrix is irreducible and the Markov chain has a unique stationary distribution. It is readily verified that the probability distribution

$$\pi_j := \frac{d_j}{\sum_{k\in\mathcal{N}} d_k}, \quad j \in \mathcal{N}, \tag{1}$$

satisfies the detailed balance equations $\pi_j p_{jk} = \pi_k p_{kj}$ for all $j, k \in \mathcal{N}$. Hence, the Markov chain (Y_n) is reversible, with unique stationary distribution π.

Let T_i denote the random time of the first return to node i for the random walk started in i, i.e., T_i is the smallest $n > 0$ for which $Y_n = i$. Recall that the counter value produced by the Random Tour algorithm is

$$X = \sum_{n=0}^{T_i-1} f(Y_n),$$

where $f(j) = \phi(j)/d_j$. In order to compute its expectation, we rely on the *cycle formula* for so-called *regenerative processes* (see [5], Chapter VI for definitions and rigorous statements). Basically, a regenerative process is a stochastic process the trajectories of which can be broken into cycles, the cycles being independent from one another and identically distributed. In the present setting, (Y_n) is a regenerative process, as can be seen by defining a cycle to start with a visit to state i, and to end before the next visit to state i. In this context, the cycle formula yields that, for an arbitrary function f:

$$\frac{\mathbf{E}_i \sum_{n=0}^{T_i-1} f(Y(n))}{\mathbf{E}_i(T_i)} = \mathbf{E}_\pi f(Y(n)) = \sum_{j\in\mathcal{N}} \pi_j f(j); \quad (2)$$

here $\mathbf{E}_\pi$ denotes expectation when the process $\{Y_n\}_{n\geq 0}$ is in a stationary regime (so that the marginal distribution of Y_n is π, for any n), $\mathbf{E}_i$ denotes the expectation when it is started from the initial state $Y_0 = i$, and T_i is the (random) time of the first n such that $Y_n = i$, i.e. T_i is the beginning of the second cycle. Specialising (2) to the indicator function $f(j) = \mathbf{1}_{j=i}$ (that is, $f(j) = 0$ unless $j = i$) yields

$$\mathbf{E}_i(T_i) = 1/\pi_i = \frac{\sum_{j\in\mathcal{N}} d_j}{d_i}. \quad (3)$$

Finally, note that the counter value X obtained in the algorithm above when the original message returns to the sending node i is exactly

$$X = \sum_{n=0}^{T_i-1} f(Y_n),$$

where $f(j) = \phi(j)/d_j$. We thus have by formula (2) that

$$\begin{aligned}\mathbf{E}_i(X) &= \mathbf{E}_i(T_i) \sum_{j\in\mathcal{N}} \pi_j f(j) \\ &= \frac{\sum_{k\in\mathcal{N}} d_k}{d_i} \sum_{j\in\mathcal{N}} \frac{d_j}{\sum_{k\in\mathcal{N}} d_k} \frac{\phi(j)}{d_j} \\ &= \frac{\Phi}{d_i}.\end{aligned}$$

Since $\widehat{\Phi} = d_i X$, this last expression guarantees that the mathematical expectation of $\widehat{\Phi}$ is indeed equal to Φ. □

3.3 Analysis of Variance

Here we specialise the discussion to the case where $\phi \equiv 1$, that is we consider only estimation of the system size N. We shall give bounds on the variance in this case, which involve a global parameter of the graph, namely the *spectral gap* of the graph. Before providing a definition, we re-interpret our estimation procedure in terms of *Continuous Time* Random Walks (CTRW).

Consider the CTRW defined as follows. After entering a node j, the walker stays there for exactly $1/d_j$ time units, and then moves to a randomly selected neighbour of node j. We shall denote by Y_t the position of the walker at time t. To distinguish from the discrete time case, we let τ_i denote the first time $t > 0$ when the walker moves from another node to node i. It is readily seen that our estimate Φ also reads

$$\Phi = d_i \int_0^{\tau_i} \phi(Y_t) dt, \quad (4)$$

and thus in the special case where $\phi \equiv 1$, this reads $\Phi = d_i \tau_i$.

We now introduce some more notation required to describe bounds on the variance of Φ.

Definition 1 The Laplacian matrix of a graph G is by definition the matrix L such that $L_{ij} = -1$ if $i \neq j$, and (i, j) is an edge of the graph G, $L_{ij} = d_i$ if $j = i$, and $L_{ij} = 0$ otherwise. Its eigenvalues $\lambda_1, \ldots, \lambda_N$ are real, non-negative. Assuming they are sorted in non-decreasing order ($\lambda_1 \leq \lambda_2 \ldots \leq \lambda_N$), then $\lambda_1 = 0$, and λ_2 is called the spectral gap of the graph.

This Laplacian matrix is intimately connected to another CTRW defined on the set of nodes of the graph, namely the CTRW where a walker's visit to a given node i lasts for an exponentially distributed duration, with mean $1/d_j$. This is the standard CTRW on a graph; it is a continuous time Markov process, and the matrix $-L$ is its *infinitesimal generator*. The CTRW we described previously differs from the standard one because sojourn times are deterministic, rather than exponentially distributed.

We now state our result concerning the variance of the estimator Φ.

Proposition 2 *For an arbitrary undirected graph on N nodes with spectral gap λ_2, the Random Tour estimate Φ of the number of nodes N verifies:*

$$N^2\left[2(1-1/N)^2-1\right]-Nd_i \leq \mathit{Var}\,(\Phi) \leq N^2\frac{2d_i}{\lambda_2}. \quad (5)$$

The proof is deferred to the appendix. We now interpret this result. Consider the lower bound first. When d_i is small compared to N (which one expects to be the normal situation) and N is large, the lower bound on the variance of Φ is equivalent to N^2. Thus, the standard deviation $\sqrt{\mathrm{Var}(\Phi)}$ of the estimate is at least of the order of its mean, that is N. As for the upper bound, it is then of order $N^2[2d_i/\lambda_2]$. Provided d_i/λ_2 is of order 1, this matches the order of the lower bound, and the estimate Φ has a standard deviation precisely of the order of its mean.

3.4 The role of graph expansion

To illustrate this further, introduce the notation

$$I(G) := \inf_{S:|S|\leq N/2}\frac{E(S,\overline{S})}{|S|},$$

where $E(S,\overline{S})$ denotes the number of edges in the graph G between the set of nodes S, and the complementary set $\overline{S}$. The constant $I(G)$ is known as the isoperimetric constant of the graph, or also as its conductance; see e.g. Mohar [31] for further discussion. The so-called Cheeger inequality (see [31]) states that the spectral gap λ_2 of the graph verifies

$$\lambda_2 \geq \frac{I(G)^2}{2\Delta(G)},$$

where $\Delta(G)$ is the maximal degree of nodes in the graph. Combined with (5), Cheeger's inequality entails that

$$\mathrm{Var}\,(\Phi) \leq N^2\frac{4d_i\Delta(G)}{I(G)^2} \leq N^2\left(\frac{2\Delta(G)}{I(G)}\right)^2. \quad (6)$$

This illustrates how the ratio $\Delta(G)/I(G)$ controls the quality of the estimator Φ. The conductance parameter I is sometimes called the expansion parameter of a graph, and graphs with large I are referred to as expanders. The reader can find additional material on expanders in [3], or [27]. Several overlay architectures proposed in the literature ensure good expansion properties by design: the expansion parameter I is bounded away from 0.

In particular, overlays comprising sufficiently many "random" edges have large expansion parameter. It is shown for instance in [17], Theorem 5.4, that Erdős-Rényi graphs on N nodes with average degree d such that $d >> \log(N)$ have an expansion of $d/2$. It is also shown in [18] that, if each node chooses $m \geq 2$ other nodes uniformly at random as its neighbours in the overlay, then the resulting graph has expansion at least $m/5$.

3.5 Complexity/Accuracy trade-off

We measure the complexity of a single Random Tour by the number of (discrete-time) steps taken by the random walk during that tour, that is T_i with the above notation. Thus, in view of (3), k consecutive Random Tours launched by node i cost on average $k\sum_{j\in\mathcal{N}} d_j/d_i$. Denote the estimates of the corresponding Random Tours by $\Phi(1),\ldots,\Phi(k)$. Their empirical mean $\overline{\Phi}$ has a variance k times smaller than that of an individual estimate. By Tchebitchev's inequality, we obtain that for a given $\epsilon > 0$, the relative error $|\overline{\Phi}-N|/N$ verifies:

$$\mathbf{P}(|\overline{\Phi}-N|/N \geq \epsilon) \leq \frac{\mathrm{Var}\overline{\Phi}}{N^2\epsilon^2}.$$

If we can tolerate a relative error greater than ϵ with probability of p_e for some $p_e \in (0,1)$, then in view of (6) and the previous display it suffices to take

$$k \geq \left(\frac{2\Delta(G)}{\epsilon I(G)}\right)^2 \frac{1}{p_e}.$$

Assuming for concreteness $p_e = 10\%$ is fixed, and the graph is *regular*, that is all its nodes have degree d, then for k runs we get on average a cost of kN, and we can guarantee a relative error of

$$\epsilon = \frac{2d}{I(G)} \frac{1}{\sqrt{kp_e}}.$$

Thus the cost is linear in N, and an extra cost factor of k buys a reduction in relative error of order $1/\sqrt{k}$.

4 The Sample&Collide method

In this section we present an algorithm which is based on, and improves upon a technique proposed in [7]. This technique essentially relies on sampling uniformly at random from the peer population. It then uses such random samples to produce an estimate of system size, based on how many random samples are required before two samples return the same peer.

We improve the proposal of [7] in two ways. First, we propose a uniform peer sampling technique which produces unbiased samples by emulating a *continuous time* random walk (CTRW), contrary to existing proposals which rely on *discrete time* random walks (DTRW) and suffer from a bias whenever peers have unequal degrees.

Second, we refine the way those samples are used, and effectively obtain estimates with a given variance with fewer sampling steps.

4.1 Peer sampling with CTRW

The probing peer's label is again denoted by i, and the overlay is modelled as an undirected graph G. For peer sampling, we shall use the standard CTRW, namely the random walk where each visit to a node j lasts for an exponentially distributed, random time with expected duration $1/d_j$, where d_j is the degree of node j. The stationary distribution of the standard DTRW puts mass $d_j / \sum_k d_k$ on each node j, and is thus biased towards high degree nodes. In contrast, the CTRW we just described has a uniform stationary distribution. Our peer sampling algorithm proceeds as follows.

1. A timer is set at some predefined value $T > 0$, by the initiator, node i, in a sampling message.
2. Any node j, either after receiving the sampling message, or (if it is the initiator) after having initialised the timer, does the following. It picks a random number U, uniformly distributed on $[0,1]$. It decrements T by $-\log(U)/d_i$ ($T \leftarrow T + \log(U)/d_i$). If $T \leq 0$, then this node j is the sampled node; it returns its ID to the initiator, and the procedure stops. Otherwise, it forwards the message with the updated timer to one of its d_j neighbors, chosen uniformly at random.

This procedure returns a random node sample, the distribution of which is exactly that of the state of the standard CTRW at time T, started from node i. This follows from the well-known fact that $-\log(U)$ has a unit mean exponential distribution. See Ross [35], and in particular the *inverse function method* for random numbers generation, for a proof.

A convenient measure of accuracy of the proposed sampling technique is provided by the *variation distance* between the probability distribution of the returned sample, and the target uniform probability distribution. Recall that the variation distance between two measures p, q on $\mathcal{N}$ is defined as

$$d(p,q) := \frac{1}{2} \sum_{i \in \mathcal{N}} |p_i - q_i|.$$

It admits the following useful interpretation [26, Theorem 5.2]: a random sample X from a distribution p coincides, with probability $1 - d(p,q)$, with a random sample from distribution q. We can now relate the quality of our

sampling method to the choice of T, and the spectral gap of the graph:

Lemma 1 *Let $\{X_t\}_{t\geq 0}$ be a continuous time, reversible Markov process on a finite state space $\mathcal{N}$, with spectral gap λ_2, and stationary distribution π. Denote by $p_{i\cdot}(t)$ the distribution of X_t when the process is started at $X_0 = i$. Then it holds that*

$$d(p_{i\cdot}(t), \pi) \leq \frac{1}{2\sqrt{\pi_i}} e^{-\lambda_2 t}.$$

The proof is given in the appendix. When specialised to the standard CTRW, for which $\pi_i = 1/N$, taking $T = c\log(N)/\lambda_2$, this reads

$$d(p_{i\cdot}(T), \pi) \leq \frac{1}{2} N^{(1/2)-c}.$$

If for instance $c \geq 3/2$, then this variation distance is of order N^{-1}.

Thus, in view of the above-mentioned interpretation of variation distance, for such a choice of T, X_T coincides with a uniform random sample from $\mathcal{N}$ with probability $1 - O(N^{-1})$. Equivalenty, it takes on average of the order of N samples before retrieving an improperly selected node.

The reason this is important is the following: we shall see that our algorithm requires of the order of $\sqrt{N}$ uniform random samples in order to estimate N. But since we cannot sample uniformly at random exactly, we use the CTRW procedure described above to obtain *approximately uniform* random samples. The error estimate on the approximation tells us that for T chosen as above, it is *as if* we sampled exactly from the uniform distribution; with high probability, our sampling procedure will yield the same samples (on runs of length $o(N)$) as the exact procedure.

We should mention that, as both N and λ_2 are a priori unknown, it is in practice not feasible to set T to precisely say, $2\log(N)/\lambda_2$. One possibility is to use sampling with a first value of T, get back from the Sample&Collide procedure (described in next Subsection) an estimate $\widehat{N}_1$ of N, then re-run the whole procedure with $2T$ instead of T, get a new estimate $\widehat{N}_2$ of T, and repeat until estimates $\widehat{N}_i$ appear to stabilise; they should increase with T until T is sufficiently large.

Another possibility is to assume suitable lower bounds on λ_2, and upper bounds on $\log(N)$ are known, so that a conservative value of T can be used. This is the approach we take in this paper.

Remark 1 Instead of the standard CTRW, we could alternatively base sampling on the CTRW with deterministic sojourn times. This suppresses the need to generate uniform random numbers U at nodes traversed by the random walk. However, in general there is no analogue of Lemma 1 for the deterministic sojourn time CTRW, as the following counter-example shows.

Consider a bipartite, regular graph with common node degrees d, nodes being partitioned into $\mathcal{N}_1$ and $\mathcal{N}_2$, with $|\mathcal{N}_1| = |\mathcal{N}_2|$. Then with the latter CTRW, whenever $\lfloor T/d \rfloor$ is even, the returned node belongs to the same bipartition as the node i from which sampling started, no matter how large T is. Assume say that node i belongs to $\mathcal{N}_1$. Then the variation distance between the sampled distribution and the uniform distribution is at least $\sum_{j\in\mathcal{N}_2} 1/N = 1/2$, and does not go to zero.

4.2 Estimation procedure

The technique we use is as follows. We pick an integer $\ell > 0$, which will determine the accuracy of our estimates. We then obtain node samples $X(1), \ldots, X(n)$. Denote by C_1 the first time n when a sample $X(n)$ is obtained which has already been seen (the first collision), i.e., for some $m < n$, $X(m) = X(n)$. Likewise, denote by C_2 the second time n when the corresponding sample $X(n)$ has previously been observed, and define C_i similarly for $i \geq 1$. We shall stop sampling at $n = C_\ell$, that is, when exactly ℓ newly obtained samples have previously been observed, where ℓ is a fixed control parameter.

For a given N, we denote by $L_N(n_1, \ldots, n_\ell)$ the probability that $C_1 = n_1, \ldots, C_\ell = n_\ell$. Elementary com-

binatorics show that

$$L_N(n_1,\ldots,n_\ell) = \\ N^{-n_\ell}\left[N(N-1)\cdots(N-n_\ell+\ell+1)\right]\times \\ \times\left[(n_1-1)(n_2-2)\cdots(n_\ell-\ell)\right]. \quad (7)$$

This formula can be written as the product of two terms, one that is a function of $(n_1,\ldots,n_\ell)$ only, and another that is a function of N and n_ℓ only. This implies that C_ℓ is a *sufficient statistic* for the estimation of N: all the information about the unknown parameter N that is carried by the observations $C_1,\ldots,C_\ell$ is contained in the variable C_ℓ. Or to put it another way, given C_ℓ, the other C_i's don't contain any additional information about N. Hence, the best estimator (for any performance measure) based on $C_1,\ldots,C_\ell$ is a function of C_ℓ only.

Our approach to estimating N will be to use the Maximum Likelihood (ML) method. Note that

$$\frac{\partial}{\partial N}\log L_N(C_1,\ldots,C_\ell) = -\frac{C_\ell}{N} + \sum_{i=0}^{C_\ell-\ell-1}\frac{1}{N-i} \\ = \frac{1}{N}\Big[-\ell + \sum_{i=0}^{C_\ell-\ell-1}\frac{i}{N-i}\Big] \quad (8)$$

The derivative of the log-likelihood function is called the score, and the expectation of its square is called the Fisher information. These quantities play a role in determining the variance of the optimal estimator, as we shall see later.

It is clear from the formula above that the likelihood is well defined for N in the interval $N \in (C_\ell-\ell-1,+\infty)$, is increasing on the interval $(C_\ell-\ell-1,\widehat{N}]$, and decreasing on $[\widehat{N},+\infty)$, where $\widehat{N}$ is the ML estimate. (It is also clear that C_ℓ can't be bigger than $N+\ell$.) Thus the ML estimate $\widehat{N}$ can be readily computed by solving the equation

$$F(N) := \sum_{i=0}^{C_\ell-\ell-1}\frac{i}{N-i} - \ell = 0, \quad (9)$$

using standard bisection search. Before detailing the procedure, we note the following. The monotonic decreasing function $F(N)$ satisfies

$$\frac{(C_\ell-\ell-1)(C_\ell-\ell)}{2N} - \ell \le F(N) \\ \le \frac{(C_\ell-\ell-1)(C_\ell-\ell)}{2(N-(C_\ell-\ell-1))} - \ell.$$

Each of the bounding functions is also monotonic decreasing in N. This readily implies that the ML estimate $\widehat{N}$ lies in the interval $[N^-,N^+]$, where

$$\begin{cases} N^- = \frac{(C_\ell-\ell-1)(C_\ell-\ell)}{2\ell}, \\ N^+ = \frac{(C_\ell-\ell-1)(C_\ell-\ell)}{2\ell} + C_\ell - \ell - 1. \end{cases} \quad (10)$$

The binary search determination of $\widehat{N}$ then proceeds as follows. Initialize the search range $[N^-,N^+]$ with the values given in (10). Then repeat the following step until $N^+-N^- \le 1$. Set $N=(N^++N^-)/2$; if $F(N)>0$, set $N^-=N$; otherwise set $N^+=N$.

4.3 Accuracy/Complexity trade-off

We now provide an asymptotic analysis of the quality of the proposed ML estimate when N is large. This will then be used to analyse the accuracy/complexity trade-off of the Sample&Collide procedure, under the assumption that samples returned by the CTRW module are indeed uniformly distributed.

Proposition 3 *Let $\ell>0$ be fixed. As N tends to infinity, we have the following convergence in distribution:*

$$\frac{C_\ell^2}{2N} \to E_1+\cdots+E_\ell, \quad (11)$$

where $E_1,\ldots,E_\ell$ are i.i.d. random variables, that are exponentially distributed with parameter 1.

Furthermore, for any positive p, we also have convergence of the p-th moments:

$$\mathbf{E}_N\left[\left(\frac{C_\ell^2}{2N}\right)^p\right] \to \mathbf{E}\left[(E_1+\cdots+E_\ell)^p\right]. \quad (12)$$

Proof: We prove the weak convergence property by induction on ℓ. We shall evaluate the following conditional probability:

$$\mathbf{P}_N(C_\ell - C_{\ell-1} > b\sqrt{N} | C_{\ell-1} = a_N\sqrt{N}),$$

where $a_N \sim a$, a is a fixed positive number, as $N \to \infty$. Let $m = a_N\sqrt{N} - (\ell - 1)$, and $k = b\sqrt{N}$. Elementary combinatorics show that this conditional probability equals

$$\frac{1}{N^k}(N-m)(N-m-1)\cdots(N-m-k+1),$$

and hence:

$$\begin{aligned}
&\log\Big(\mathbf{P}_N\Big(C_\ell - C_{\ell-1} > b\sqrt{N}\Big|\, C_{\ell-1} = a_N\sqrt{N}\Big)\Big)\\
&= \sum_{i=0}^{k-1}\log\Big(\frac{N-m-i}{N}\Big) \sim -\sum_{i=0}^{k-1}\frac{i+m}{N}\\
&\sim -ab - \frac{b^2}{2}.
\end{aligned}$$

We thus have the following equivalent as $N \to \infty$:

$$\mathbf{P}_N\Big(C_\ell - C_{\ell-1} > b\sqrt{N}\;\Big|\; C_{\ell-1} = a_N\sqrt{N}\Big) \sim e^{-ab-b^2/2}.$$

In turn, setting $a = \sqrt{2y}$ and $b = \sqrt{2(x+y)} - \sqrt{2y}$, we get

$$\mathbf{P}\Big(\frac{C_\ell^2}{2N} > x + y \;\Big|\; \frac{C_{\ell-1}^2}{2N} = y\Big) \sim e^{-ab-b^2/2} = e^{-x}.$$

This establishes the claimed weak convergence property.

In order to deduce convergence of moments from weak convergence, it is enough to show that for all $p > 0$, the distributions of variables $[C_\ell/\sqrt{N}]^p$ for varying N are uniformly integrable (for a definition see e.g. [8]). By a standard criterion for uniform integrability, this will follow if for some positive θ, it holds that

$$\sup_{N>0}\mathbf{E}_N\left[e^{\theta C_\ell/\sqrt{N}}\right] < \infty. \tag{13}$$

By a simple coupling argument (see [26] for background on coupling), it can be shown that the distribution of C_ℓ is stochastically dominated by that of the sum of ℓ independent copies of C_1. Thus,

$$\mathbf{E}_N\left[e^{\theta C_\ell/\sqrt{N}}\right] \le \left[\mathbf{E}_N\left(e^{\theta C_1/\sqrt{N}}\right)\right]^\ell.$$

It is therefore enough to prove (13) in the special case where $\ell = 1$. Write

$$\begin{aligned}
\mathbf{E}_N\left[e^{\theta C_1/\sqrt{N}}\right] &= \int_0^\infty \mathbf{P}_N\left(e^{\theta C_1/\sqrt{N}} > y\right)dy\\
&= \int_0^\infty \mathbf{P}_N\left(C_1 > \frac{\sqrt{N}}{\theta}\log(y)\right)dy.
\end{aligned}$$

We bound the integrand in the last expression as follows. Set $k = \sqrt{N}\log(y)/\theta$. Then,

$$\begin{aligned}
\mathbf{P}_N(C_1 > k) &= \exp\Big(\sum_{i=0}^{k-1}\log(1 - i/n)\Big)\\
&\le \exp(-(k-1)^2/N).
\end{aligned}$$

Combined with the previous expression, this yields

$$\begin{aligned}
&\mathbf{E}_N\left[e^{\theta C_1/\sqrt{N}}\right]\\
&\le 1 + \int_1^\infty \exp\Big(-\frac{(\sqrt{N}\log(y)/\theta - 1)^2}{N}\Big)dy\\
&= 1 + \int_{-1/\sqrt{N}}^\infty e^{-v^2}e^{\theta(v+1/\sqrt{N})}\theta dv\\
&\le 1 + \int_{-1}^\infty e^{-v^2+\theta v+\theta}\theta dv,
\end{aligned}$$

where the equality is obtained by the change of variables $v = \log(y)/\theta - 1/\sqrt{N}$. The final term is finite and independent of N, which implies the announced uniform integrability. □

This proposition allows us to evaluate the asymptotic mean square error of the ML estimate:

Corollary 1 *The ML estimate $\widehat{N}$ is such that*

$$\lim_{N\to\infty}\frac{1}{N^2}\mathbf{E}_N\left(\widehat{N} - N\right)^2 = \frac{1}{\ell}. \tag{14}$$

The proof of this corollary is deferred to the Appendix. It shows that, for a variance of the order of N^2/ℓ, we use C_ℓ samples, hence on average a number of samples of the order of $\sqrt{N\ell}$. The average number of messages exchanged in a single sampling step is, assuming the originator is randomly selected, equal to $T\overline{d}$, where $\overline{d} = N^{-1}\sum_j d_j$ is the average node degree. Thus, assuming as in Section 4.1 that T equals $2\log(N)/\lambda_2$, the average number of messages used by the Sample&Collide method is of order $(\overline{d}\log(N)\sqrt{N\ell}/\lambda_2)$, for an estimate with relative variance of $1/\ell$. This presents an improvement on the cost of the "inverted birthday paradox method" of [7] by a factor $\sqrt{\ell}$.

Next, we show that no other unbiased estimator can do substantially better.

Lemma 2 *Let $\tilde{N} = f(C_\ell)$ be an arbitrary estimator with the property that $\mathbf{E}_N[f(C_\ell)] = N$. Then,*

$$\liminf_{N\to\infty} \frac{1}{N^2} Var_N(\tilde{N}) \geq \frac{1}{\ell}.$$

The proof is in the appendix. It is a well-known result in statistics that the maximum likelihood estimator is asymptotically efficient (has asymptotic mean square error no larger than that of the best unbiased estimator), but this is in the context where the parameter to be estimated is fixed, while the number of samples increases to infinity. In the setting we are studying, both the parameter and the number of observations go to infinity; hence, we have included a proof of optimality.

Remark 2 Observe that, since $C_\ell^2/(2\ell) \sim N$ (in probability and expectation), the bounds N^- and N^+ in (10) are both asymptotic to N, and differ only by a term of order $\sqrt{N}$. Hence, instead of computing the maximum likelihood estimator by binary search, we could equally well use either N^- or N^+, or the asymptotically unbiased estimator $\tilde{N} = C_\ell^2/(2\ell)$. All three estimators are within $\sqrt{N}$ of the ML-estimator, and hence, all three are asymptotically efficient. In fact, for ease of computation, we use the estimator $\tilde{N} = C_\ell^2/2\ell$ in the next section, where we evaluate the algorithm.

We now compare the accuracy/cost trade-offs of our two methods. Assume for ease of discussion that degrees are constant, equal to d. Recall that for the Random Tour method, for a cost of order kN, we have an upper bound on the relative variance of $2d/(k\lambda_2)$. In order to match the variances of the two methods, we need to set $k = 2\ell d/\lambda_2$. Thus the ratio of costs of the methods becomes

$$\frac{\text{cost}(RT)}{\text{cost}(S\&C)} = \frac{(2\ell d/\lambda_2)N}{2d\log(N)\sqrt{N\ell}/\lambda_2} = \frac{\sqrt{N\ell}}{\log(N)}.$$

Thus for large systems (large N), or for accurate measurements (large ℓ), the Sample&Collide method should be preferred.

5 Experiments

We first describe the setting of the experiments. We then report results on the accuracy and cost of the two methods in a static environment, and finally consider dynamic environments with both gradual and abrupt changes in system sizes.

5.1 Setup and evaluation criteria

Our experiments are simulation-based. We consider overlay networks of exactly 100,000 nodes in the static case, and comprising between 50,000 and 150,000 nodes in the dynamic case. We consider two classes of topologies in the evaluation, which we refer to as balanced and scale-free random graphs.

Overlays of the first type (balanced random graphs) are generated so as to guarantee node degrees lying between 1 and 10, in the following manner. Sequentially, each node i selects a random number d_i^{out} between 1 and 10. It then selects d_i^{out} target nodes at random, among target nodes with a current degree less than 10. Then d_i^{out} undirected edges are created between node i and its d_i^{out} targets, whose degree is increased by 1 at this stage. The resulting average degree is between 7 and 8. From the results of [18], we expect such graphs to have large expansion, hence a favourable situation for our two techniques. Existing overlay maintenance protocols aim to maintain graphs with similar statistical properties; see e.g. [22] and [16].

In scale-free networks on the other hand, the node degree distribution follows a power law; there is evidence that the Internet and the World-Wide-Web have this property. We generate random scale-free graphs using the preferential attachment scheme of Barabási and Albert [1]. Here, each new node added to the network chooses its links prefentially targeting high-degree nodes. The result is a random graph in which the probability that a node has k neighbors decays like k^{-3}. Thus, node

degrees are much more variable than in the balanced random graph model.

In the dynamic scenarios, newly incorporated nodes are connected via their own set of random targets, chosen according to the rule for the corresponding model. Nodes to be removed are selected uniformly at random, and the remaining nodes that lose neighbors do not search for new ones. The actual system size we report is always that of the connected component to which the probing node belongs.

We evaluate our algorithms for estimating system size on the following metrics. *Accuracy* relates to the relative error in the system size estimate and is clearly a basic criterion. It can be improved by taking more measurements. So there is a tradeoff with the *Overhead*, specified as the number of messages required to obtain the system size estimate. Depending on the application, a quick approximate estimate could be preferable to a more accurate one which would take much longer to compute, and create more overhead. This could also be the case when churn is high, causing the system size to change rapidly. In that case, *Reactivity to changes* is an important characteristic of the algorithms. To evalutate this, we compute the time to react to a growth or increase in the number of peers in the system.

5.2 Results in static settings

5.2.1 Balanced random graphs

We did repeated runs on a 100,000 node overlay of the following schemes: Random Tour (RT), Sample&Collide (S&C) with $\ell = 10$, and S&C with $\ell = 100$. For both instances of S&C, the timer value used in the sampling module was fixed to $T = 10$. In view of our suggestion to take $T \approx 2\log(N)/\lambda_2$, this is consistent with a spectral gap λ_2 larger than 2.3.

Figure 1 displays the empirical average of estimates obtained by RT, as a percentage of the actual value, ranging from one to 3000 estimates. The cost increases linearly with the number of runs, and the variance of the averaged estimate decreases like the reciprocal of this number. The three different curves correspond to three distinct generated graphs on which the measurements were launched. The curves plotted in Figure 2 display

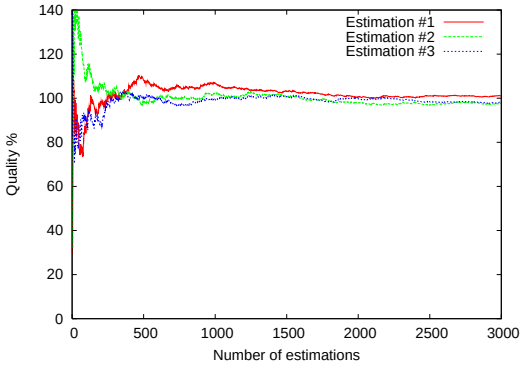

Fig. 1 Empirical averages of Random Tour estimate values (as percentage of system size) over increasing numbers of samples, on a 100,000 node overlay random graph.

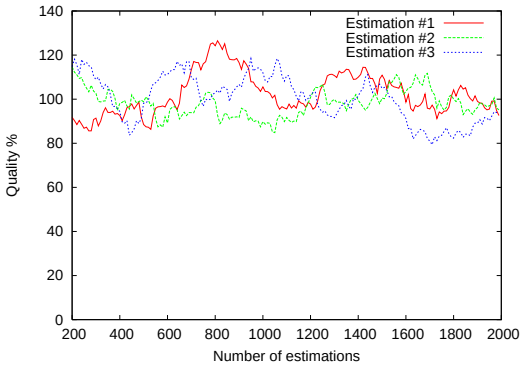

Fig. 2 Empirical averages of Random Tour esimate values (as percentage of system size) on a sliding window of the last 200 samples, on a 100,000 node overlay random graph.

estimates obtained by taking the empirical average on a sliding window of 200 samples. This choice of sliding window size corresponds to a standard deviation of 0.2, roughly consistent with an accuracy of +/-20%.

Figure 3 plots a run of S&C with $\ell = 100$. It shows that S&C with $\ell = 100$ needs about an order of magnitude fewer samples to achieve the same accuracy as RT, which is consistent with the theoretical analysis.

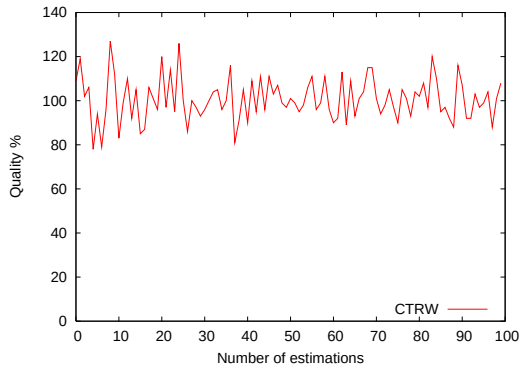

Fig. 3 Sample and collide with $l = 100$ (no sliding window), on a 100,000 node random graph.

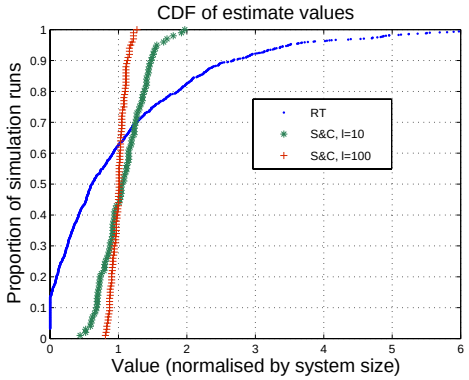

Fig. 4 CDF of estimate values, normalised by system size, on a 100,000 overlay random graph, for Random Tour, Sample&Collide and with $\ell = 10$ and $\ell = 100$.

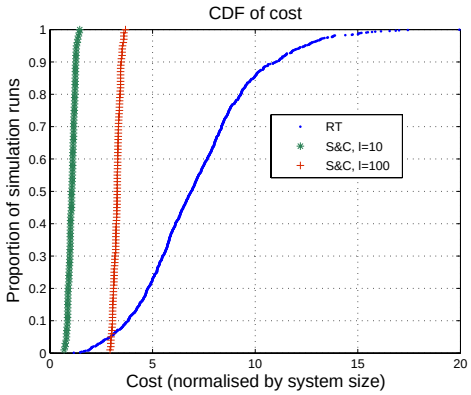

Fig. 5 CDF of estimation cost in messages, normalised by system size, on a 100,000 overlay random graph, for Random Tour, and Sample&Collide with $\ell = 10$ and $\ell = 100$.

The cumulative distribution functions (cdfs) of the normalised estimate values for the three candidate methods, RT, S&C (ℓ=10) and S&C (ℓ=100) are displayed on Figure 4. The steeper the curve, the less dispersed the sample values. This is further illustrated by the summary statistics reported in Table 5.2.1. All three methods provide samples with the correct mean value; RT has a larger variance than S&C (ℓ=10), which in turn has a larger variance than S&C (ℓ=100). Note how the variances of both S&C methods match the theoretical prediction, and coincide with $1/\ell$.

We next report on the costs incurred in a single run of each of the three methods. The cdfs of costs, normalised by system size, are shown in Figure 5. It is clear that the costs of S&C are far less variable than those of RT. The mean and variance of normalised cost, reported in Table 5.2.1, show that the cost of RT is both higher and more variable. Note how S&C(ℓ=100) incurs a cost per run that is larger than that of S&C(ℓ=10) by only a factor of 3.27 (consistent with the ratio of $\sqrt{100}/\sqrt{10} \approx 3.16$ predicted by the analysis), for a variance reduction by a factor of 10.

Algorithm	RT	S&C, $\ell = 10$	S&C, $\ell = 100$
Average value	1.01	1.08	1.01
Variance(value)	1.3	0.1	0.01
Average cost	7.16	1.08	3.27
Variance(cost)	8.06	0.1	0.02

Table 1 Summary statistics of sampling strategies: mean and variance of normalised estimate values, and mean and variance of normalised estimate costs.

5.2.2 Scale-free graphs

Figures 6 and 7 depict the system size estimates as a percentage of the actual system size, on scale-free graphs in the static scenario. The plots show that both the Random Tour and Sample&Collide methods achieve accuracy comparable to what they achieved in the balanced random graph setting. This suggests that they are capable of dealing with considerable node heterogeneity in providing unbiased estimates of system size.

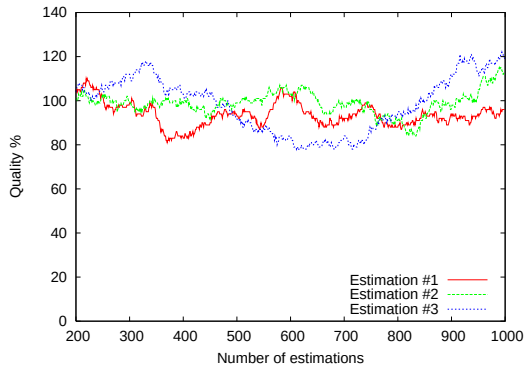

Fig. 6 Random Tour with sliding window (last 200), on a 100,000 node scale-free graph.

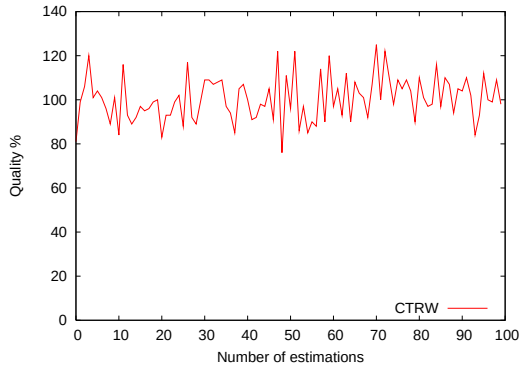

Fig. 7 Sample and collide with $l = 100$ (no sliding window), on a 100,000 node scale-free graph.

5.3 Results in dynamic settings

We did repeated runs on random graphs with varying numbers of nodes of the two following procedures: Random Tour (RT), and Sample&Collide (S&C) with $\ell = 10$ and $\ell = 100$. For space reasons we don't show the results for $\ell = 10$.

We considered three distinct dynamic scenarios:

- Gradual decrease, where the node population steadily decreases from 100,000 to 50,000;
- Gradual increase, where the node population grows regularly from 100,000 to 150,000;
- Catastrophic changes, where the initially node population of 100,000 is suddenly decreased to 75,000 and then to 50,000, and finally faces a flash crowd with a sudden arrival of 25,000 nodes.

For both estimation procedures, we consider using sliding windows over past sampled values. The size of the sliding window conditions both the accuracy, with a variance estimate proportional to the reciprocal of the window size, and the reactivity of the estimates, which are less reactive to changes for larger windows. For S&C with $\ell = 100$, we choose not to average, i.e. take a sliding window of size 1.

5.3.1 Random Tour

The performance of RT is illustrated on a shrinking network (gradual decrease scenario) in Figure 8, on a growing network (gradual increase scenario) in Figure 9, and on a network with catastrophic failures and flash crowd arrivals in Figure 10. A sliding window of 700 samples is used in all three cases.

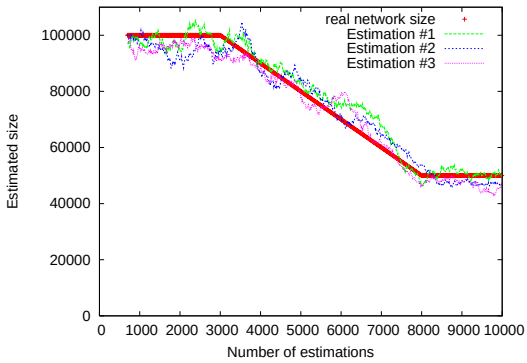

Fig. 8 Random Tour with sliding window (last 700) on shrinking network; 100,000 nodes at beginning, 50% nodes removal from run 3000 to run 8000.

The key observation to make from these figures is the following. The overlay structure can be drastically affected by population changes; in particular, node departures may reduce the expansion parameter of the initial overlay, which would lead to a poorer accuracy of the RT estimators. Nevertheless, in all three scenarios, we observe that RT maintains a constant accuracy level throughout the changes in system size. In summary, RT is robust to changes in system sizes.

We also note, particularly in Figure 10, an increased convergence time of the estimate due to the averaging sliding window; obviously, the smaller the window, the faster the convergence time but the higher the estimator variance.

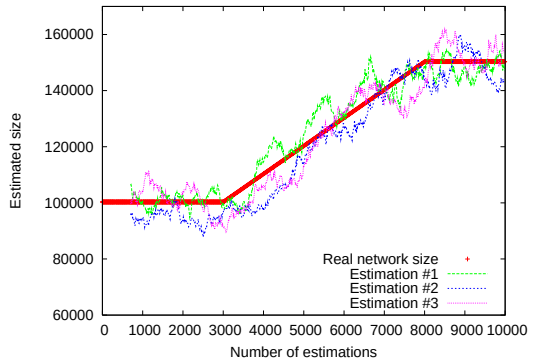

Fig. 9 Random Tour with sliding window (last 700) on growing network; 100,000 nodes at beginning, 50% nodes join from run 3000 to run 8000.

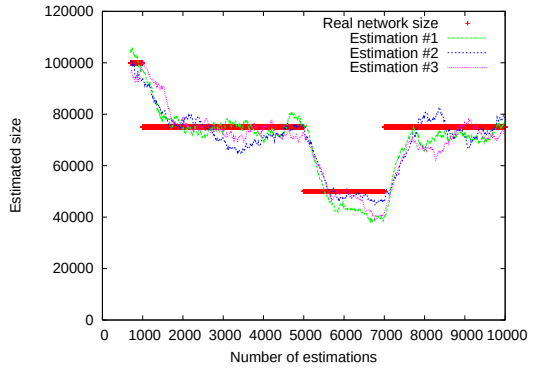

Fig. 10 Random Tour with sliding window (last 700) on network with catastrophic failures failures; 100,000 nodes at beginning, -25,000 nodes at runs 1000 and 5000, +25,000 nodes at run 7000.

We should mention at this stage that in the simulations reported here, we did not allow a departing node to leave the system with the probing message. Such a situation may however arise in practice, for instance due to node crash or improper departure from the overlay.

One way of handling such message loss at the node initiating the measurement is to declare a probing message to be lost if it has not been recovered in a given duration since its launch. The corresponding timeout parameter needs to be sufficiently large so that only few Random Tour packets time out while the corresponding message is not lost, and still traveling through the system. One could for instance set this time-out to the average trip time, plus a few multiples of the trip time standard deviation. Here trip time refers to real-world time, as measured by the initiator's clock. Both standard deviation and average of trip time can be estimated adaptively from past trip time measurements.

A similar solution could be applied to protect the sampling procedure used in S&C against packet losses; now the trip time is the time till the walk stops, rather than the time till the walker returns to the originator.

5.3.2 Sample&Collide

The performance of S&C is illustrated on a shrinking network (gradual decrease scenario) in Figure 11, on a growing network (gradual increase scenario) in Figure 12, and on a network with catastrophic failures and flash crowd arrivals in Figure 13.

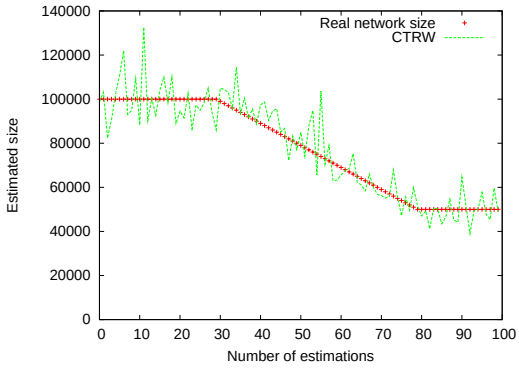

Fig. 11 Sample&Collide with $\ell = 100$, no sliding window, on shrinking network; 100,000 nodes at beginning, 50% nodes removal from run 30 to run 80.

Changes in system size could affect the performance of S&C, because the quality of the random samples returned by the sampling module deteriorates if the expan-

sion of the overlay is reduced. However, as we observe on the three figures, the S&C estimator maintains a constant level of accuracy. The analysis predicts a relative variance of $1/\ell = 1/100$ for individual estimates, under the assumption of perfect random sampling. Hence theory predicts a relative standard deviation of 10% for the estimates plotted on these figures, provided sampling works well. The fluctuations on figures 11, 12 and 13 are consistent with a 10% magnitude.

Thus we conclude that S&C, and its sampling module, are robust to changes in system size, both gradual and sudden. We emphasize here that the costs incurred are much lower than for RT; a single point on Figure 11 costs on average 350,000 messages; a single point on Figure 8 costs on average 560 millions messages, i.e., three orders of magnitude larger.

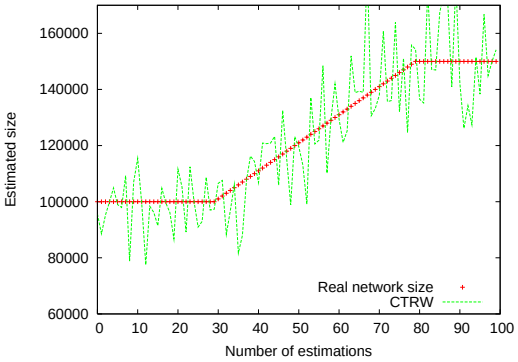

Fig. 12 Sample&Collide with $\ell = 100$, no sliding window, on growing network; 100,000 nodes at beginning, 50% nodes join from run 30 to run 80.

6 Conclusion

We addressed the issue of estimating the size of large-scale peer-to-peer overlay networks. We proposed two peer counting approaches based on random walks. More generally, these approaches may be used to count the number of peers with given characteristics or to estimate aggregate statistics. This is useful for basic overlay

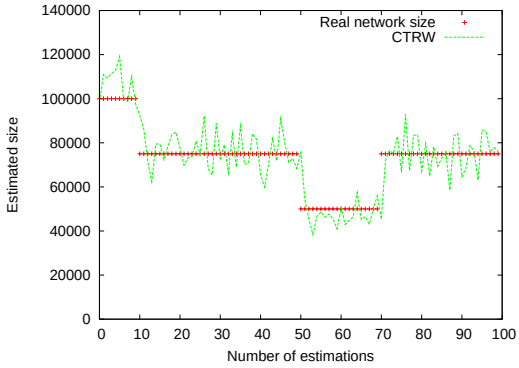

Fig. 13 Sample&Collide with $\ell = 100$,no sliding window, under catastrophic failures: -25,000 nodes at runs 10 and 50, +25,000 nodes at run 70.

maintenance, and we expect it to be useful as well for applications such as live media streaming.

The Random Tour method aggregates local node statistics along a "random tour", that is a random walk stopped when it returns to its starting node. It is simple to implement, and is suitable for small to moderate systems. Its cost scales linearly with system size, like several other proposals we reviewed, such as gossip or random polling. It does not incur the ACK implosion problem that random polling techniques face. We analysed the properties of the Random Tour estimator, and showed that it is unbiased, and how its standard deviation is controlled by the expansion parameter of the overlay.

The Sample&Collide method requires random samples of peers. We proposed a peer sampling algorithm based on a Continuous Time Random Walk and showed that it produces asymptotically uniform samples, in contrast to previous proposals which were biased towards high degree nodes. We showed that its cost for a specified accuracy is characterised again by the expansion parameter of the overlay. We constructed a system size estimate based on the number of samples required to observe duplicated samples. We analysed in detail the asymptotic properties of this estimate, and showed that it makes the most efficient use of the information in the samples, by achieving the smallest possible variance. To our knowledge this achieves the best cost / accuracy trade-off of

proposals to date[30], with a cost scaling like the square root of the system size, and the square root of the required accuracy (measured in reciprocal of relative variance). It is therefore a suitable candidate for large scale environments. Random Tour can still be attractive in moderately large systems, when one is interested in measuring aggregates of node properties rather than just system size.

Finally we evaluated our two schemes via simulations, in both static and dynamic environments. The simulation results confirmed the theoretical analysis of the two schemes. We found that both were robust to system changes, both gradual and sudden.

References

1. R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics 74, 47*, 2002.
2. D. Aldous and J. Fill. Reversible Markov chains and random walks on graphs. Mnograph in preparation, available at *http://stat-www.berkeley.edu/users/aldous/book.html*.
3. N. Alon and J. Spencer. *The probabilistic method*. Wiley, 2002.
4. S. Alouf, E. Altman, and P. Nain. Optimal on-line estimation of the size of a dynamic multicast group. In *Proc. IEEE INFOCOM*, 2002.
5. S. Asmussen. Applied probability and queues. *Springer*, 2003.
6. F. Baccelli and P. Brémaud. *Elements of Queueing Theory*. Springer, 2003.
7. M. Bawa, H. Garcia-Molina, A. Gionis, and R. Motwani. Estimating aggregates on a peer-to-peer network. *Technical Report, Dept. of Computer Science, Stanford University*.
8. P. Billingsley. *Convergence of probability measures*. Wiley, 1999.
9. J.-C. Bolot, T. Turletti, and I. Wakeman. Scalable feedback control for multicast video distribution in the internet. In *Proc. ACM SIGCOMM*, 1994.
10. S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah. Gossip algorithms: Design, analysis and applications. In *Proc. IEEE INFOCOM*, 2005.
11. M. Castro, P. Druschel, A. Ganesh, A. Rowstron, and D. wallach. Security for structured peer-to-peer overlay networks. In *Proc. OSDI*, 2002.
12. T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, 1991.
13. D. Dolev, O. Mokryn, and Y. Shavitt. On multicast trees: structure and size estimation. In *Proc. IEEE INFOCOM*, 2003.
14. P. Eugster, S. Handurukande, R. Guerraoui, A.-M. Kermarrec, and P. Kouznetsov. Lightweight probabilistic broadcast. *ACM Trans. Computer Systems, 21(4)*, November 2003.
15. T. Friedman and D. Towsley. Multicast session membership size estimation. In *Proc. IEEE INFOCOM*, 1999.
16. A. J. Ganesh, A.M. Kermarrec, and L. Massoulié. Peer-to-peer membership management for gossip-based protocols. *IEEE Trans. Computers, vol.52, no.2*, 2003.
17. A.J. Ganesh, L. Massoulié, and D. Towsley. The effect of network topology on the spread of epidemics. In *Proc. IEEE INFOCOM*, 2005.
18. A.J. Ganesh and F. Xue. Expansion properties of k-out random graphs. Preprint, 2004.
19. K. Horowitz and D.Malkhi. Estimating network size from local information. *Information Processing Letters 88(5):237–243*, 2003.
20. M. Jelasity and A. Montresor. Epidemic-style proactive aggregation in large overlay networks. In *Proc. ICDCS*, 2004.
21. M. Jelasity and M. Preuß. On obtaining global information in a peer-to-peer fully distributed environment. *Springer LNCS, Vol.2400, pp. 573-577*.
22. M. Jelasity, S. Voulgaris, R. Guerraoui, and A-M. Kermarrec. Gossip-based peer sampling. submitted.
23. A.-M. Kermarrec, L. Massoulié, and A. J. Ganesh. Efficient application-level multicast on a network-aware self-organizing overlay. *IRISA internal publication, 1166-8687;1657*, 2004.
24. D. Kostoulas, D. Psaltoulis, I. Gupta, K. Birman, and A. Demers. Decentralized schemes for size estimation in large and dynamic groups. *IEEE NCA '05*, 2005.
25. J. Li and D.-Y. Lim. A robust aggregation tree on distributed hash tables. *MIT Student Oxygen Workshop*, 2004.
26. T. Lindvall. *Lectures on the Coupling Method*. Dover, 2002.
27. N. Linial and A. Wigderson. Expander graphs and their applications. Available at http://www.cs.huji.ac.il/ nati/.
28. D. Malkhi, M. Naor, and D. Ratajczak. Viceroy: a scalable and dynamic emulation of the butterfly. In *Proc. ACM Symp. PODC*, 2002.
29. G. S. Manku. Routing networks for distributed hash tables. In *Proc. ACM Symp. PODC*, 2003.
30. E. Le Merrer, A.-M. Kermarrec, and L. Massoulié. Peer-to-peer size estimation in large and dynamic networks: a comparative study. In *Proc. 15th IEEE Intl. Symp. High Perf. Distr. Comp. (HPDC 15)*, 2006.

31. B. Mohar. Some applications of Laplace eigenvalues of graphs, in *Graph symmetry: algebric methods and applications*, G. Hahn and G. Sabidussi eds., NATO ASI Ser. C 497, Kluwer, 1997, pp. 225-275.
32. J. Nonnenmacher and E. W. Biersack. Optimal multicast feedback. In *Proc. IEEE INFOCOMM*, 1998.
33. D. Psaltoulis, D. Kostoulas, I. Gupta, K. Birman, and A. Demers. Practical algorithms for size estimation in large and dynamic groups. In *Proc. ACM Symp. PODC*, 2004.
34. Dimitrios Psaltoulis. Private communication. July 2005.
35. S. Ross. *Simulation.* Elsevier, 2001.
36. X. Zhang, J. Liu, B. Li, and T.-S. P. Yum. Donet / coolstreaming: A data-driven overlay network for live media streaming. In *Proc. IEEE INFOCOM*, 2005.

7 Appendix

7.1 Proof of Proposition 2

We shall in fact derive a more general result. To this end, we need to introduce notation, and a key result on reversible Markov processes described in detail in [2].

Consider a continuous time Markov process $\{X_t\}$ on a finite set $\mathcal{N}$. Let $\{\pi_i\}_{i\in\mathcal{N}}$ denote its stationary distribution, assumed to be unique. Let Q be the infinitesimal generator of the Markov process $\{X_t\}$. Denote the diagonal term q_{ii} by $-q_i$. In particular, each visit to a given state i lasts for an exponentially distributed random time with mean q_i^{-1}. Assume the process is reversible, and denote by λ_2 the spectral gap of the generator, i.e. $-\lambda_2$ is the second largest eigenvalue of Q. Let $\mathbf{E}_{stat}$ denote expectation when the initial state of the process is distributed according to the stationary distribution π. Finally, let T_i^0 denote the first time $t>0$ such that $X_t=i$. One then has:

Lemma 3 (Aldous-Fill [2], Ch. 3 p.21) *For any state $i\in\mathcal{N}$, one has*

$$\frac{(1-\pi_i)^2}{q_i\pi_i} \le \mathbf{E}_{stat}(T_i^0) \le \frac{1-\pi_i}{\lambda_2\pi_i}. \tag{15}$$

Given a Markov process $\{X_t\}$ as above, consider the process $\{X'_t\}$ which has the same state space $\mathcal{N}$, visits the states in the same order as the other process, but whose visits to each state i last for a deterministic duration q_i^{-1}, instead of an exponentially distributed random duration. We denote by T_i the first *entrance* time of X'_t into state i, that is the smallest $t>0$ such that $t^-\neq i$ and $t^+=i$. The previous lemma will be used to establish the following

Proposition 4 *For any process $\{X'_t\}$ as above, started at the beginning of a visit to state i, it holds that*

$$\begin{array}{l}\frac{2(1-\pi_i)^2-1}{(\pi_i q_i)^2}+\frac{2}{\pi_i q_i}\left[\frac{1}{q_i}-\sum_{j\in\mathcal{N}}\frac{\pi_j}{2q_j}\right] \le Var(T_i)\\ \le \frac{1}{(\pi_i q_i)^2}\left[\frac{2q_i(1-\pi_i)}{\lambda_2}-1\right]+\frac{2}{\pi_i q_i^2}.\end{array} \tag{16}$$

Proof: The process $\{X'_t\}$ is regenerative, with as regeneration points the entrances into state i. By the cycle formula, it holds that

$$\mathbf{E}_{stat}(T_i) = \frac{\mathbf{E}_i\int_0^{T_i}(T_i-t)dt}{\mathbf{E}_i(T_i)} = \frac{\mathbf{E}_i(T_i^2)}{2\mathbf{E}_i(T_i)}.$$

It thus follows that

$$\mathrm{Var}(T_i) = 2\mathbf{E}_i(T_i)\mathbf{E}_{stat}(T_i) - \left(\mathbf{E}_i(T_i)\right)^2. \tag{17}$$

We shall now relate $\mathbf{E}_{stat}(T_i)$ to $\mathbf{E}_{stat}(T_i^0)$, which will allow us to use the previous lemma. One has the following identity:

$$\mathbf{E}_{stat}(T_i) = \mathbf{E}_{stat}(T_i^0) + \pi_i\mathbf{E}_i(T_i) - \sum_{j\in\mathcal{N}}\frac{\pi_j}{2q_j}. \tag{18}$$

Indeed, this follows directly from the following identities.

$$\begin{cases}\mathbf{E}_{stat}(T_i) = \sum_{j\in\mathcal{N}}\pi_j\mathbf{E}_{stat}(T_i|X'_0=j),\\ \mathbf{E}_{stat}(T_i^0) = \sum_{j\neq i}\pi_j\mathbf{E}_{stat}(T_i^0|X_0=j),\\ \mathbf{E}_{stat}(T_i|X'_0=i) = \mathbf{E}_i(T_i)-\frac{1}{2q_i},\\ \mathbf{E}_{stat}(T_i|X'_0=j) = \mathbf{E}_{stat}(T_i^0|X_0=j)-\frac{1}{2q_j},\ \ j\neq i.\end{cases}$$

The first two identities are obvious. We omit the details of the proof of the last two identities for brevity; a formal argument can be constructed by making use of Palm calculus (see e.g. [6]).

By the cycle formula, $\mathbf{E}_i(T_i)=1/(\pi_i q_i)$. Combined with (17) and (18), this yields

$$\mathrm{Var}(T_i) = \frac{2}{\pi_i q_i}\left[\mathbf{E}_{stat}(T_i^0)+\frac{1}{q_i}-\sum_j\frac{\pi_j}{2q_j}\right]-\frac{1}{(\pi_i q_i)^2}.$$

The proposition follows by combining this expression with (16). □

Upon specialising Proposition 4 to the CTRW on a graph G, for which $\pi_j \equiv 1/N$, and $q_j \equiv d_j$, one easily shows that

$$N^2 \frac{2(1-1/N)^2 - 1}{d_i^2} - \frac{N}{d_i} \leq \mathrm{Var}(T_i) \leq \frac{N^2}{d_i^2} \frac{2d_i}{\lambda_2}.$$

Proposition 2 follows, as $\mathrm{Var}(\Phi) = d_i^2 \mathrm{Var}(T_i)$.

7.2 Proof of Lemma 1

By Lemma 8, page 10, Chapter 4 in [2], it holds that

$$d(p_{i\cdot}(t), \pi) \leq \frac{1}{2\sqrt{\pi_i}} \sqrt{p_{ii}(2t) - \pi_i}. \tag{19}$$

Besides, as mentioned on Eq. (46), page 20, Chapter 3 in [2], the function $t \to p_{ii}(t) - \pi_i$ is *completely monotone* (see p.19, chapter 3 in [2] for a definition), and thus, by lemma 13, p. 20, ch.3 [2], it verifies

$$p_{ii}(t) - \pi_i \leq [p_{ii}(0) - \pi_i] e^{-\lambda_2 t}.$$

Combined with (19), this yields the claim of the lemma.

7.3 Proof of Corollary 1

We have by the Cauchy-Schwarz inequality that

$$\begin{aligned}
&\mathbf{E}_N[(\widehat{N} - N)^2] \\
&= \mathbf{E}_N[(N^- - N)^2] + 2\mathbf{E}_N[(\widehat{N} - N^-)(N^- - N)] \\
&\quad + \mathbf{E}_N[(\widehat{N} - N^-)^2] \\
&\leq \mathbf{E}_N[(N^- - N)^2] \\
&\quad + 2\sqrt{\mathbf{E}_N(\widehat{N} - N^-)^2} \sqrt{\mathbf{E}_N(N^- - N)^2} \\
&\quad + \mathbf{E}_N[(\widehat{N} - N^-)^2].
\end{aligned} \tag{20}$$

In view of the bounds (10), $\widehat{N} - N^-$ is bounded in absolute value by $N^+ - N^- \leq C_\ell$. Hence,

$$\mathbf{E}_N[(\widehat{N} - N^-)^2] \leq \mathbf{E}_N(C_\ell^2) \sim 2\ell N, \tag{21}$$

by (12). Convergence of moments (12) also guarantees, in view of the expression for N^-, that

$$\mathbf{E}_N[N^-] \sim \mathbf{E}_N\Big(\frac{C_\ell^2}{2\ell}\Big) \sim N, \tag{22}$$

while

$$\begin{aligned}
\mathbf{E}_N[(N^-)^2] &\sim \mathbf{E}_N\Big(\frac{C_\ell^4}{4\ell^2}\Big) \\
&\sim \frac{N^2}{\ell^2} \mathbf{E}[(E_1 + \cdots + E_\ell)^2] \\
&= \frac{N^2}{\ell^2}(\ell^2 + \ell).
\end{aligned} \tag{23}$$

Here, we have used the fact the exponential distribution with parameter 1 has mean 1 and variance 1. Now, by (22) and (23),

$$\mathbf{E}_N[(N^- - N)^2] \sim \mathbf{E}_N[(N^-)^2] - N^2 \sim \frac{N^2}{\ell}. \tag{24}$$

Substituting (21) and (24) in (20), we get

$$\lim_{N\to\infty} \frac{1}{N^2} \mathbf{E}_N(\widehat{N} - N)^2 = \frac{1}{\ell},$$

as claimed. □

7.4 Proof of Lemma 2

Let $\tilde{N} = f(C_1, \ldots, C_\ell)$ be any unbiased estimator of N, i.e., $\mathbf{E}_N[f(C_1, \ldots, C_\ell)] = N$. We shall use the Cramér-Rao inequality (see, e.g., [12, Theorem 12.11.1]) to obtain a lower bound on the variance of this estimator. To use this inequality, we need to compute the Fisher information (a measure of the 'information' that the random vector $(C_1, \ldots, C_\ell)$ contains about the parameter N). The Fisher information $I(N)$ is defined as the variance of the score function,

$$s(N) = \frac{\partial}{\partial N} \log L_N(C_1, \ldots, C_\ell).$$

Therefore, it follows from (8) that

$$\begin{aligned}
I(N) = \frac{1}{N^2} \mathbf{E}_N\Big[\Big(\sum_{i=0}^{C_\ell - \ell - 1} \frac{i}{N-i}\Big)^2\Big] + \frac{\ell^2}{N^2} \\
- \frac{2\ell}{N^2} \mathbf{E}_N\Big[\sum_{i=0}^{C_\ell - \ell - 1} \frac{i}{N-i}\Big].
\end{aligned} \tag{25}$$

Recall that the mean of the score function is zero (see, e.g., [12, Section 12.11]). Thus, by (8),

$$\mathbf{E}_N\Big[\sum_{i=0}^{C_\ell - \ell - 1} \frac{i}{N-i}\Big] = \ell.$$

Substituting this in (25) yields

$$I(N) = \frac{1}{N^2}\mathbf{E}_N\Big[\Big(\sum_{i=0}^{C_\ell-\ell-1}\frac{i}{N-i}\Big)^2\Big] - \frac{\ell^2}{N^2}. \tag{26}$$

Now observe using (12) and Markov's inequality that, for any fixed $\epsilon > 0$

$$\mathbf{P}_N(C_\ell > \epsilon N) \le (\epsilon N)^{-6}\mathbf{E}_N[C_\ell^6] \le cN^{-3}, \tag{27}$$

for some constant $c > 0$ that depends on ϵ but not on N. Now, on the event that $C_\ell > \epsilon N$,

$$\sum_{i=0}^{C_\ell-\ell-1}\frac{i}{N-i} \le \sum_{i=0}^{N-1}\frac{i}{N-i} \le N\log N,$$

whereas, on the event that $C_\ell \le \epsilon N$,

$$\sum_{i=0}^{C_\ell-\ell-1}\frac{i}{N-i} \le \frac{1}{1-\epsilon}\frac{C_\ell^2}{2N}.$$

Hence,

$$\mathbf{E}_N\Big[\Big(\sum_{i=0}^{C_\ell-\ell-1}\frac{i}{N-i}\Big)^2\Big] \le (N\log N)^2\mathbf{P}_N(C_\ell > \epsilon N) + \frac{1}{(1-\epsilon)^2}\mathbf{E}_N\Big[\Big(\frac{C_\ell^2}{2N}\Big)^2\Big] \tag{28}$$

Now, the first term in the last expression above goes to zero as $N \to \infty$ by (27), while the second term goes to $(\ell^2+\ell)/(1-\epsilon)^2$ by (12), and well-known properties of the exponential distribution. Hence, we have from (26) and (28) that

$$I(N) \le \frac{1}{N^2}\Big(\frac{\ell^2+\ell}{(1-\epsilon)^2} - \ell^2\Big).$$

Hence, by the Cramér-Rao bound [12, Theorem 12.11.1], for any unbiased estimator $\tilde{N} = f(C_1, \ldots, C_\ell)$, we have

$$\mathrm{Var}(\tilde{N}) \ge \frac{1}{I(N)} \ge \frac{(1-\epsilon)^2N^2}{\ell + (2\epsilon-\epsilon^2)\ell^2}.$$

Since ℓ is fixed, letting ϵ decrease to zero yields the claim of the lemma.