A Comparative Study of Spanning Tree and Gossip Protocols for Aggregation*

Lehel Nyers
University of Szeged, Hungary, and Subotica Tech, Subotica, Serbia

Márk Jelasity
MTA-SZTE Research Group on Artificial Intelligence,
and University of Szeged, Hungary

Abstract

Distributed aggregation queries like average and sum can be implemented in different paradigms like gossip and hierarchical approaches. In the literature, these two paradigms are routinely associated with stereotypes such as "trees are fragile and complicated" and "gossip is slow and expensive". However, a closer look reveals that these statements are not backed up by systematic studies. A fair and informative comparison is clearly needed. However, this is a hard task because the performance of protocols from the two paradigms depends on different subtleties of the environment and the implementation of the protocols. We tackle this problem by carefully designing the comparison study. We use state-of-the-art algorithms and propose the problem of monitoring the network size in the presence of churn as the ideal problem for comparing very different paradigms for global aggregation. Our simulation study helps us identify the most important factors that differentiate between gossip and spanning tree aggregation: the time needed to compute a truly global output, the properties of the underlying topology, and sensitivity to dynamism. We demonstrate the effect of these factors in different practical topologies and scenarios. Our results help us to choose the right protocol in the light of the topology and dynamism patterns.

1 Introduction

Fully distributed aggregation is an important problem where we wish to execute queries such as sum, average, minimum, or maximum over unreliable networks (sensor networks, physical networks of routers, overlay networks, etc.), in which no central servers are directly accessible.

At least two paradigms are known that solve this problem. The first one is the *gossip* approach where algorithms were proposed to achieve large degrees of robustness. Gossip protocols do not rely on fixed topologies: nodes exchange information with random neighbors to implement a diffusion-like computation pattern, and as a result the system converges to a state where all the nodes know the query result. From the literature, here we just focus on the adaptive approaches. In [1], the authors propose the restarting technique to convert any one-shot algorithm into an adaptive one. Apart from restarting, other approaches have been proposed that focus on error correction through some form of bookkeeping at the nodes [2–5].

The second paradigm is *hierarchical aggregation*, which is a popular method in sensor networks [6]. It was also proposed for general process groups [7]. Tree-based aggregation remained unpopular in some areas like peer-to-peer networks due to the widely held assumptions about its lack of robustness. There are a few notable exceptions however: the Astrolabe framework [8], which is in fact only a virtual tree with completely unstructured gossip communication patterns behind it; the GAP protocol and its variants [5, 9–11] that actually build a spanning tree over a distributed network; and PRISM [12], a hierarchical approach that is built on top of a distributed hashtable, with a focus on detecting and signaling imprecise output.

*This is the pre-peer reviewed version of the following article: Concurrency Computat.: Pract. Exper. 2015; 27:4091–4106, which has been published in final form at http://dx.doi.org/10.1002/cpe.3549.

Unfortunately, the literature is strongly influenced by stereotypes about both approaches like "spanning tree protocols are fragile" and "gossip protocols are slow and expensive". It is tacitly assumed that these statements have been conclusively settled. However, when surveying the literature, this does not turn out to be the case. In fact, we are unaware of any studies with a focus on a careful comparison among very different paradigms for aggregation. For example, Merrer et al [13] compare algorithms for size estimation, but they do not include spanning tree methods that are in the focus of our interest. Chitnis et al consider a very basic tree protocol that has no capabilities for reconfiguration, and briefly compare it with gossip [14]. The environment they consider is sensor networks. Due to the limited scope and suboptimal representatives of gossip and tree protocols, this study does not settle the issues we raised. Wuhib et al [5] propose an adaptive gossip protocol and compare it with GAP. Interestingly, GAP outperformed the gossip protocol in all scenarios investigated (which were inspired by aggregation tasks in wired networks of routers). While this is a very nice result, it is not completely conclusive due to their selection of the communication topology and the aggregation problem.

Our main contribution is that we propose a careful empirical methodology to shed light on the strengths and weaknesses of both approaches. We simulate competitive, state-of-the-art representatives of spanning tree and gossip protocols, and model different network environments. We identify the key aspects that determine the performance of the protocols in order to help application developers select the best solution in a given practical setting.

This study is a significantly extended and improved version of our previous conference publication [15].

2 The protocols employed in our comparison study

In our system model, we assume that there are N nodes that form a network with the help of reliable channels such as TCP connections or physical links. Nodes communicate only by exchanging messages over these channels. Messages can be delayed. In addition, nodes can join and leave at any time. We assume the existence of a local failure detector at each node that sends a message to the local processes when a neighbor node fails. Leaving nodes and crashed nodes are treated identically. Leaving nodes can join again, and while offline, they retain their state. When they join again, they reconnect to their previous neighbors.

The common problem all of the following algorithms handle is the monitoring of aggregate values. That is, at any point in time t we have $N(t) \leq N$ online nodes, all of which have a value. Now let the set of values at time t be $A(t) = \{a_1(t), \ldots, a_{N(t)}(t)\}$. The task is to continuously calculate (monitor) a global function $f(A(t))$. A given algorithm for solving this problem typically supports a well-defined set of aggregate functions f.

We describe the key ideas behind all the protocols we examine in our simulation study, along with comments about our own implementation, where applicable. Our full implementation of all the protocols we used in the simulation study can also be downloaded.[1]

2.1 GAP (General Aggregation Protocol)

GAP is an adaptation of the classical self-stabilizing BFS construction algorithm of Dolev et al [16] that is based on message passing instead of shared tables. We implemented the version of GAP described in [9].

In GAP, there is a special node that acts as the root of the spanning tree. The root is fixed and guaranteed to remain available. The tree grows from the root as all the nodes discover their shortest path towards the root, starting with the neighbors of the root, and so on. GAP implicitly assumes a relatively stable underlying network. Each node in the network maintains a table that contains an entry for each neighbor and the node itself. Each table entry contains the level in the tree, and classifies the neighbor as parent, child, or peer. The parent of each node is always the neighbor with the minimal level (say, ℓ), and the node's own level is always $\ell + 1$. A table entry also contains the aggregate value in the subtree rooted in the neighbor. These values are used to calculate the node's own aggregate.

A node gets several types of messages related to changes in the topology (failed or new neighbors) or changes in the aggregate value (locally or in a subtree of a child node). When receiving a message, the

[1] `http://peersim.sourceforge.net/`

node updates its own tables if necessary in such a way that the invariants of the tree structure and aggregate calculation are restored. Our implementation uses the "cache-like" policy [9] for maintaining the table, which means that table entries change only due to explicit messages and never due to predictions.

GAP can be implemented in a reactive or a proactive manner. In the former case, all changes are immediately reported to the neighbors. In the latter case, changes accumulate during a time period and are reported at once in a round-based fashion. We implemented the proactive round-based version, as it has better load balancing and generates fewer messages on average in dynamic environments.

The original publication of GAP did not mention that it is also important that the connections with neighbors need to preserve the order of the messages, otherwise inconsistent states can occur. This can be achieved with an appropriate transport layer, or at the application level as well.

2.2 Adaptive gossip protocols

We used the push-sum algorithm as a starting point [17]. In this algorithm (as in all gossip variants) the basic idea is that the nodes engage in a diffusive computation, during which nodes periodically send to each neighbor a proportion of the "mass" they store and also receive mass from neighbors. This way the nodes can collectively compute the average of all the values. Other aggregates, such as the network size can also be computed: if a single node has a value of 1, and all the other nodes have a value of 0 then the average is $1/N$, which can be used to recover the network size N.

The push-sum algorithm is by default a one-shot algorithm, unsuitable for monitoring. There are two approaches to achieve adaptivity. The first is the restart-based approach and the second is what we call the "bookkeeping" approach. We included in our set of algorithms a representative of both classes. In both cases, in each round a node with k neighbors sends one kth of its mass to each of the neighbors.

Restarted push-sum The key idea is that the algorithm is run in *epochs* of some fixed length, after which the gossip protocol is restarted automatically in a distributed way [1]. In effect, the restart mechanism takes a snapshot of the system at the beginning of the epoch that involves the nodes that were live at that time, and then the aggregate of this snapshot is computed during the epoch. After the completion of the epoch, the computed aggregate value is used as the output of the algorithm, hence the output is delayed by roughly two epoch lengths at most. Depending on the topology of the network, epochs can be rather short (as few as 20 rounds) due to the quick convergence of gossip.

LiMoSense, a bookkeeping approach Instead of restarting, a gossip protocol might attempt to repair the state of the nodes as a reaction to failure. This can be achieved if some variant of bookkeeping for the underlying gossip algorithm (e.g. push-sum) is implemented that makes it possible to "undo" those computations that had to do with a failed node, or that makes it possible to repair message drop failure by comparing books with neighbors. The main design goal of such protocols is the classical requirement of self-stabilization, that is, to be able to eventually converge after failures and dynamism stop. A state-of-the-art representative of such protocols is LiMoSense [2]. We use this protocol in our comparison study.

2.3 Common properties

When comparing different paradigms, we should focus on application areas and systems where the paradigms being investigated are all feasible and have a similar cost. In other words, there are systems that are *obviously* suitable only for one or the other algorithm. For example, if the network is very reliable and static, then using a tree-based approach is obviously better: the main issue with tree-based approaches is fault tolerance, and with that issue out of the way, a tree protocol provides optimal efficiency. On the other hand, if there is not a single reliable node that can be assigned the role of the root of the tree, then gossip approaches are clearly better, since without a reliable root spanning tree protocols are much more complicated and error prone (every time the root fails, a new root has to be elected in a distributed way using a consensus protocol and the entire tree needs to adapt to the new root). Here, we attempt to avoid these obvious cases, and instead we attempt to characterize the systems in which both protocols are potentially applicable.

First, the system is assumed to have a *special stable node* that is guaranteed to remain available in the network. GAP crucially relies on such a node to act as the root of the tree for tree building and maintenance. Such a node is not critical for gossip but—given that due to GAP we need to assume a stable root—gossip protocols can and will take advantage of it too. For example, when calculating the network size, the node that has the initial value of 1 can be the root (see Section 3.1). Note that GAP does not rely on the root for reading out the value: it can be modified to propagate the global aggregate to all the nodes.

Second, all the protocols are *round based* with a period (round length) of Δ. They generate a very *similar amount of traffic* in each round: each node sends one message to each neighbor in each round. In the case of GAP this can be substantially reduced, but only when the network becomes static and there is no failure. This is because no messages need to be sent if there is no change in the aggregated value or in the underlying topology. In our implementation, GAP broadcasts in each round even if there is apparently no change. The reason is that—since we work with systems that constantly change—this results in a negligible amount of extra traffic, and it solves a subtle issue of the original algorithm related to churn.

3 Simulation Setup and Methodology

We performed our comparative study with the help of the PeerSim [18] simulator using the event-based engine.

3.1 Network Size as the Aggregation Problem of Choice

Calculating the average of distributed values is often the baseline problem used to evaluate generic distributed aggregation algorithms. This, however, is rather problematic because the performance then depends crucially on the distribution of the values. If the distribution is concentrated around the average, then one cannot differentiate between the ability of an algorithm to provide real global results and between the local sampling effect, that is, when the average of local samples is similar to the global average by pure chance. This is true in the case of both gossip and spanning tree algorithms.

It is vital that here we wish to compare the *global* behavior of the algorithms, that is, how they behave in scenarios where they need to consider the entire data set. The performance of such global tasks can be considered a *worst case*, which can only improve when local neighborhoods already offer a good approximation of a given query. Of course it is of interest to know how certain algorithms react to specific distributions, and one could even develop algorithms that explicitly exploit specific known distributions, if such prior knowledge is available. However, without prior knowledge getting a quick result due to local sampling is just a matter of chance, so when comparing very different generic paradigms, we consider a robust worst case analysis more informative and preferable.

Our choice is the network size estimation problem. For this problem, the spanning tree approach counts each node according to the tree hierarchy: all nodes have a value of 1, and the tree calculates the sum. The gossip protocols here will calculate the average in a network of N nodes where the initial value is 0 at all nodes except the root, where it is 1, which gives $1/N$ as a result [1]. In both cases, the point is that the problem is clearly global, where a useful answer is available only after the algorithm has globally converged.

3.2 Network Topologies

Our protocols need undirected topologies, so where the original topology definition is directed, it has to be understood with the directionality of the edges dropped. All networks are of size $N = 1000$ unless otherwise stated.

NewsCast A dynamic topology defined in [19]. In a nutshell, without describing NEWSCAST in detail, each node will have a new set of random neighbors in each cycle using the same cycle period as the aggregation protocol. The number of neighbors is $k = 30$. The motivation for including NewsCast is that gossip protocols are often implemented over such dynamic topologies so that nodes can communicate with random samples from the network in each cycle, as assumed in theoretical discussions of gossip protocols.

Random k-out A static topology in which every node connects to a set of k random neighbors. After dropping directionality, the average degree is $2k$. The motivation for including this topology is that randomly sampled, but static, topologies have been proposed recently as the optimal choice in commercial P2P platforms over the Internet [20].

Binary Tree An undirected balanced binary tree is formed. In our experiments the root node of the aggregation protocols is placed at different levels of the tree from 0 (the root of the binary tree) to $\lceil \log_2 N \rceil$, the leaf level of the tree. We include this artificial topology to be able to illustrate a major difference between the gossip approach and spanning tree approach.

Barabasi-Albert (BA) To test heavy tailed degree distributions, we include the BA network that is constructed incrementally. New nodes connect to old nodes already in the network according to the preferential attachment rule, that is, with a probability proportional to the degree of the old node [21]. New nodes get $k = 2$ edges when they are added to the topology. In our experiments the root node of the aggregation protocols is placed at nodes with different degrees in this topology.

3.3 Failure and Churn Scenarios

We used the same model of message delay in each experiment: each message is delayed by a uniform random time drawn from the interval $[0, 0.2\Delta]$. Our preliminary experiments revealed very little sensitivity to message delay in all the protocols, so we do not focus on this aspect. We consider no message drop failures. This is because in most scenarios that are reasonable for a spanning tree the underlying topologies in question are static, so it is feasible to apply a reliable transport layer such as TCP.

The protocols require a failure detector. We assume a timeout-based detector with a timeout of 5Δ in all the experiments. Our preliminary experiments suggested very little sensitivity to this parameter as well, so we keep it fixed throughout the study.

Node churn was modeled based on statistics from a BitTorrent trace [22] as well as known empirical findings [23]. We draw the online session length for each node independently from a log-normal probability distribution with two different parameter settings. The first setting that we call *fast churn* is $\mu = 3$ and $\sigma^2 = 1$, which results in a mean of ~ 33. The unit of the resulting online session lengths is the communication period Δ. This—considering the fact that Δ can be expected to be in the range of seconds—is a rather short session length so it represents a very dynamic scenario. The second set of parameters that we call *slow churn* is $\mu = 6$ and $\sigma^2 = 2$, which results in a mean of ~ 1096.

Offline session lengths are determined implicitly by fixing the number of nodes that are online at the same time. The ratio of online nodes was set to a range of values from $\alpha = 1$ to $\alpha = 0.2$. As stated previously, nodes that re-join the network retain the state they had when leaving the network.

3.4 Evaluation methodology and Metrics

We are interested in static behavior. As mentioned above, we assume a constant churn pattern with a static expected network size αN, where α is the ratio of online nodes in the scenario in question. In this setting, we expect a good monitoring algorithm to consistently signal $\hat{N}(t) = \alpha N$ as the approximated network size in cycle t.

To measure how close a given algorithm is to this optimal behavior, we run each scenario 10 times for 10,000 cycles, and collect statistics of the absolute error $|\alpha N - \hat{N}(t)|/(\alpha N)$ over the last 9,000 cycles for each run. We ignore the first 1,000 cycles in order to allow the system to reach an equilibrium state. We plot the average and the standard deviation (with error bars).

4 Results

First, we demonstrate some weaknesses of the gossip approaches and GAP that are not so evident at first sight. This will shed light on which scenarios to avoid for these paradigms. Subsequently, we look at the

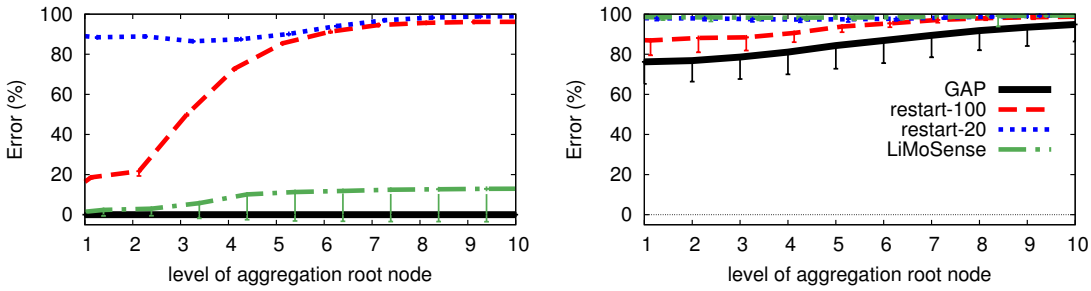

Figure 1: Binary tree, no churn (left) and fast churn, 80% of the nodes online (right). The horizontal axis represents the level of the aggregation root node in the physical topology (0: root, 10: leaves). The label restart-m refers to restarted gossip with epoch length m.

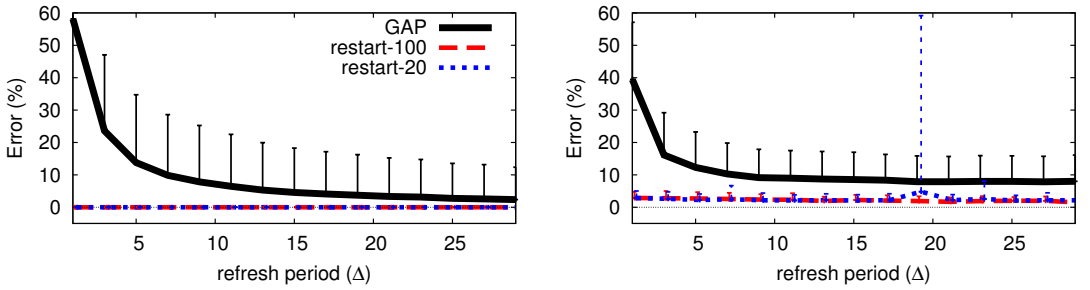

Figure 2: NewsCast with $k = 30$, no churn (left) and fast churn, 80% of the nodes online (right). The horizontal axis represents the neighborhood refresh period of NewsCast in rounds (Δ). The label restart-m refers to restarted gossip with epoch length m.

two realistic topologies: static random k-out and the BA topology, and examine some interesting subtleties that are important in these cases, and that define which approaches are preferable.

4.1 The Achilles heel of gossip

In principle, gossip protocols for aggregation have been shown to work on any connected topology, that is, they are guaranteed to converge. However, if the communication topology is not a complete graph, but instead a *static* graph with relatively small average degree, then the convergence speed is well-known to depend on the mixing time of a random walk on this topology [24]. On the other hand, the convergence time of GAP depends on the diameter (that is, the maximal minimal path length) that bounds the maximal number of steps information needs to take to reach the root.

It is often the case that graphs with a low diameter also have a rapid mixing time, but trees are exceptions. For example, the rooted balanced binary tree has a diameter of $O(\log N)$, while it has a mixing time of $O(N)$ (see, for example, [25], Example 7.7).

For this reason, at least in the failure-free scenario, we expect gossip protocols to suffer, and this is indeed the case as Figure 1 (left) shows. GAP achieves full precision very quickly, whereas even LiMoSense does not reach convergence within 10,000 rounds when the node that is assigned the role of the root node is closer to the leaves in the physical topology, let alone restart that is inherently limited in the number of rounds until convergence. (We remind the reader not to confuse the root node of the aggregation with the root of the physical topology.) However, when we introduce churn, all algorithms suffer since the underlying topology is very fragile. Still, GAP performs best (see Figure 1 (right)).

The results also reveal another important point, namely it does matter a lot where the root node is placed within the underlying physical topology. It is much harder to break out of a region closer to the leaves for the diffusion process as it is from the root (recall that for gossip the aggregation root is initialized with a

value of 1, while the remaining nodes have a value of 0).

4.2 The Achilles heel of spanning trees and bookkeeping gossip protocols

In many cases, gossip protocols assume there is a random set of neighbors in each round [1] that is given by a dynamic protocol for peer sampling [19]. This radically dynamic neighbor set is ideal for vanilla gossip. However, if bookkeeping is involved, it becomes a serious problem, since the tables will grow indefinitely until they reach the size of the whole network. This is not scalable, since all entries have an associated failure detector as well, which need to maintain a communication link with each node. For this reason, to get scalability, the members of the old neighbor set should be treated as failed nodes. This, however, ruins the ability of the protocol to converge if the aggregation task is global, as in network size estimation. All in all, with dynamic peer sampling bookkeeping gossip cannot be applied at all with any hope of success.

For GAP, the changing neighbor set raises similar issues: growing tables (and eventually a spanning tree with a star topology) or the option of treating old neighbors as failed. In our implementation, we opt for the second approach, as the option of growing tables is clearly not scalable.

Figure 2 shows simulation results with the NewsCast dynamic topology. Clearly, for fast refreshing periods the only feasible protocol is restarted gossip. Yet in the case of slower refreshing (when the topology becomes relatively stable in the short run) GAP is competitive. LiMoSense is the least favorable option in this scenario.

4.3 The k-out topology

We examined the k-out topology for different values of k. Without churn, all the protocols can achieve an error that is practically 0% for $k \geq 2$, except for restart-20 that achieves an error of 25% for $k = 2$. Clearly, an epoch length of 20 is not sufficient for such a low value of k. Note that the lower the value of k the greater the mixing time.

Figure 3 shows the results of our experiments involving churn. Our first observation is that GAP and restart are rather insensitive to the speed of churn, whereas LiMoSense is very sensitive. In the slow churn scenario the results for the latter dramatically improve. This is because fast churn is highly disruptive for this algorithm due to the constant attempts to repair the state of the system when a neighbor leaves.

At the same time, with slow churn all the algorithms become rather unstable when the offline session lengths are long (that is, when α is small). This is because—although the network is relatively stable—in such scenarios the aggregation root can become disconnected and can remain so for a relatively long time, which temporarily results in extremely large errors.

As for GAP, we observe an interesting case that is consistent with our findings over the binary tree topology: when k is very small, GAP has a slight advantage due to not depending on the mixing time of the topology. Note that for a very small k the random k-out topology behaves locally like a tree as there is a rather small probability for finding short circles, which slows gossip protocols down.

For large values of k gossip protocols can take advantage of the very good mixing properties and can beat GAP, especially with an epoch length of 100. GAP also profits from an increasing k (and therefore a decreasing diameter, and more options to repair the tree) but not as much as gossip protocols.

The error of all the protocols remains quite high. A main source of error in all the cases is the partitioning of the network. The most frequent type of partitioning in k-out networks is when isolated nodes become disconnected from the large connected cluster, as larger clusters have a much larger number of links and thus a smaller probability of losing all their connections to the main cluster [19]. The most problematic case is when the root gets disconnected, which results in a huge error with a small probability. To increase the stability of the root we repeated our simulations while keeping not only the root but also its neighbors alive. The results are shown in Figure 4. As we can see, the error is dramatically reduced, as a result of keeping the root connected to the large connected cluster. Our previous comments on the relationship of the several paradigms still hold true here.

We also experimented with larger networks: with N=10,000 and N=100,000, keeping the neighbors of the root alive as well. Figure 5 includes our results with N=10,000. We make two observations. First, now we need larger values of k to have a similar stability, which is a well-known property of k-out topolo-

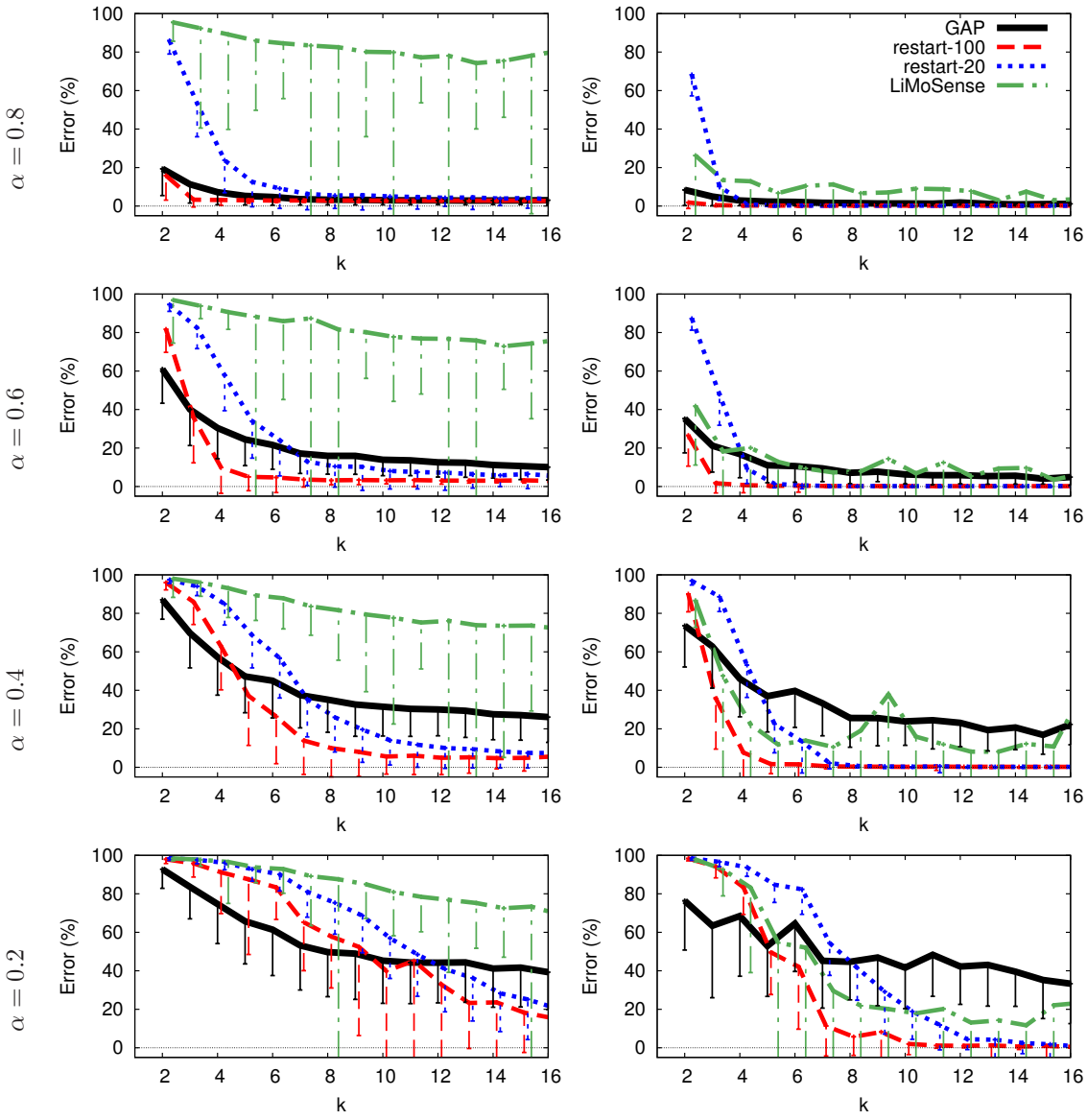

Figure 3: Error over the k-out topology, fast churn (left column) and slow churn (right column) as a function of k. The proportion of online nodes is given by α. The label restart-m refers to restarted gossip with epoch length m.

gies [19]. Second, we observe that here GAP seems to have a slightly increased error compared to that for N=1000, even when k is sufficiently large for gossip. This suggests that GAP is less robust to network size.

To verify this observation, we ran our simulations with N=100,000 and fast churn. The results are shown in Figure 6. Indeed, for all levels of churn we clearly see larger errors for GAP, while the gossip variants still converge to low levels of error, although at larger values of k.

4.4 The Barabasi-Albert topology

We generated one BA topology with $k = 2$ as previously described. In this fixed topology we placed the aggregation root at nodes with different degrees. Our results are shown in Figure 7. The range of degrees we plot include the full range of the degree distribution on a log scale.

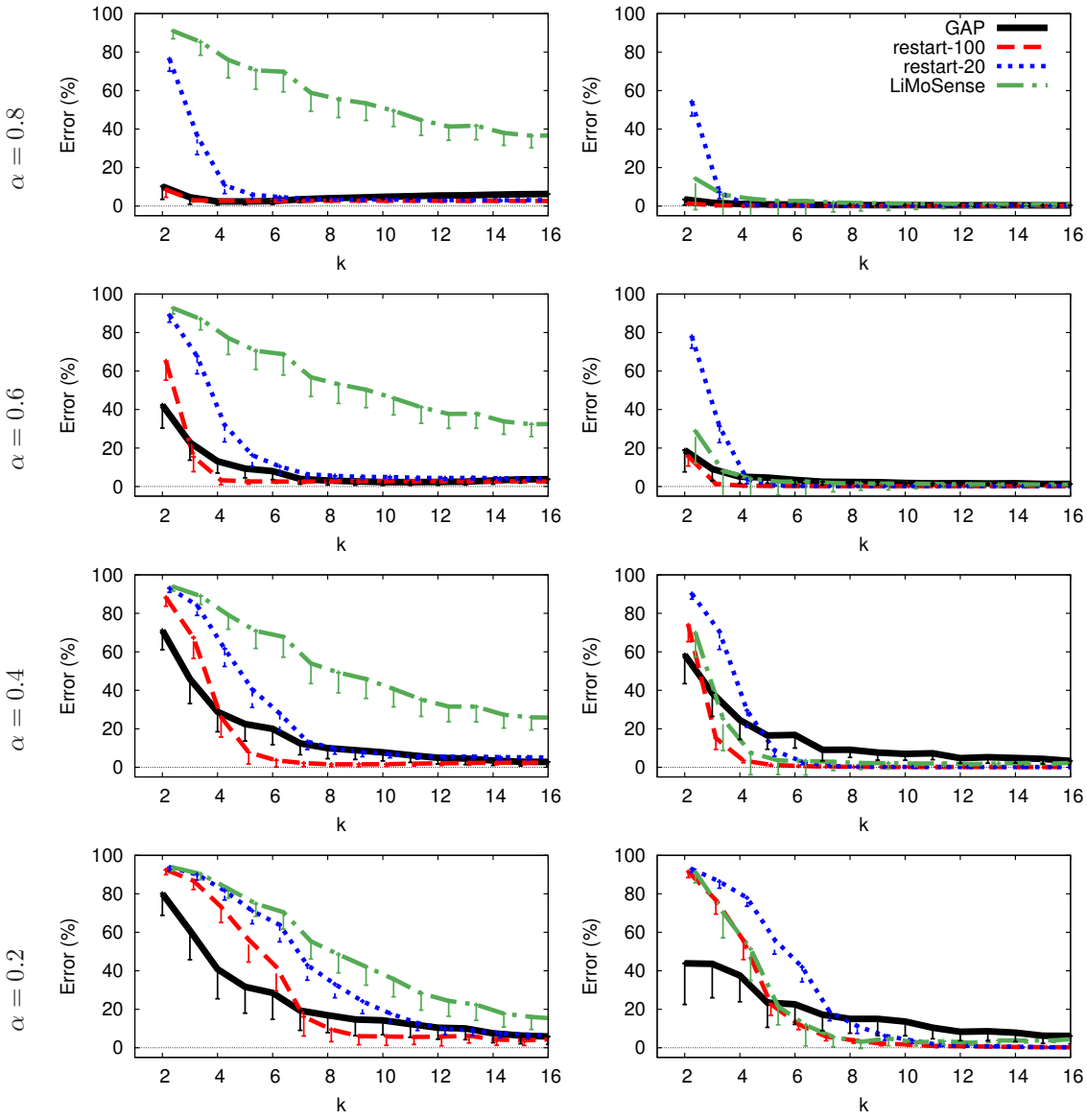

Figure 4: Error over the k-out topology with reliable root neighbors, fast churn (left column) and slow churn (right column), as a function of k. The proportion of online nodes is given by α. The label restart-m refers to restarted gossip with epoch length m.

We can observe a strong dependence of the precision of the aggregation result on the position of the aggregation root in the underlying BA topology. Placing the root on central nodes with a large degree results in a significantly lower error. Interestingly, this is also true for the gossip protocols, which here also rely on a fixed "root" node (see Section 2.3).

As in the k-out topology, without churn (not shown) all the protocols can achieve an error that is practically 0%, except restart-20 that achieves an error of 35% to 5% depending on the centrality of the root. Restart-20 performs poorly throughout the experiments, clearly indicating that an epoch length of 20 is not sufficient. At the same time, restart-100 is among the best options in most scenarios.

Clearly, in these scenarios GAP delivers the most stable performance. As with the k-out topology, gossip protocols are sensitive to the speed of churn: with LiMoSense it is more so, but the restart variants also show sensitivity, with restart-100 being the most robust.

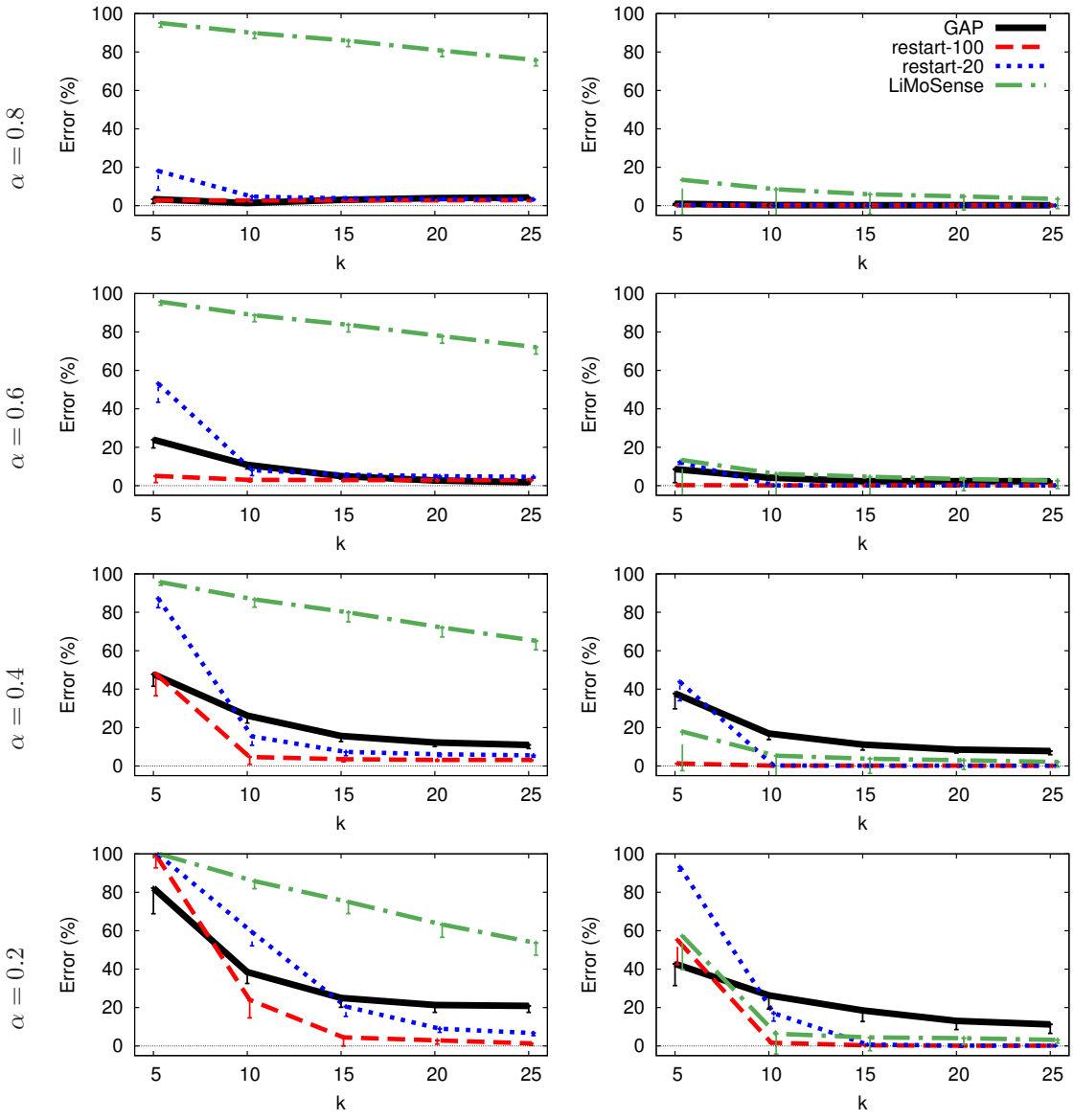

Figure 5: Error over the k-out topology with reliable root neighbors, N=10,000, fast churn (left column) and slow churn (right column), as a function of k. The proportion of online nodes is given by α. The label restart-m refers to restarted gossip with epoch length m.

As in the k-out topology, in slow churn we observe a very large variance for small α and for a low degree aggregation root. The reason is the same: the aggregation root can get disconnected. Indeed, when we simulate the scenario where the neighbors of the root are also reliable, we observe a markedly lower error for all the protocols (Figure 8).

Here, we also explore the scalability of the protocols. We experimented with network sizes of N=10,000 and N=100,000, keeping the neighbors of the root alive. In these cases not all the different node degrees were tested for root placement. Instead, we selected a representative set of nodes with varying degrees. We always included the node with the highest degree as well in this set. Figure 9 includes our results with N=10,000. Clearly, GAP is still the best alternative in all the scenarios. Overall, the error of all the protocols increases due to the larger scale. The latter observation is further supported by Figure 10 that shows our results with N=100,000 with fast churn, where the precision is further reduced due to scale.

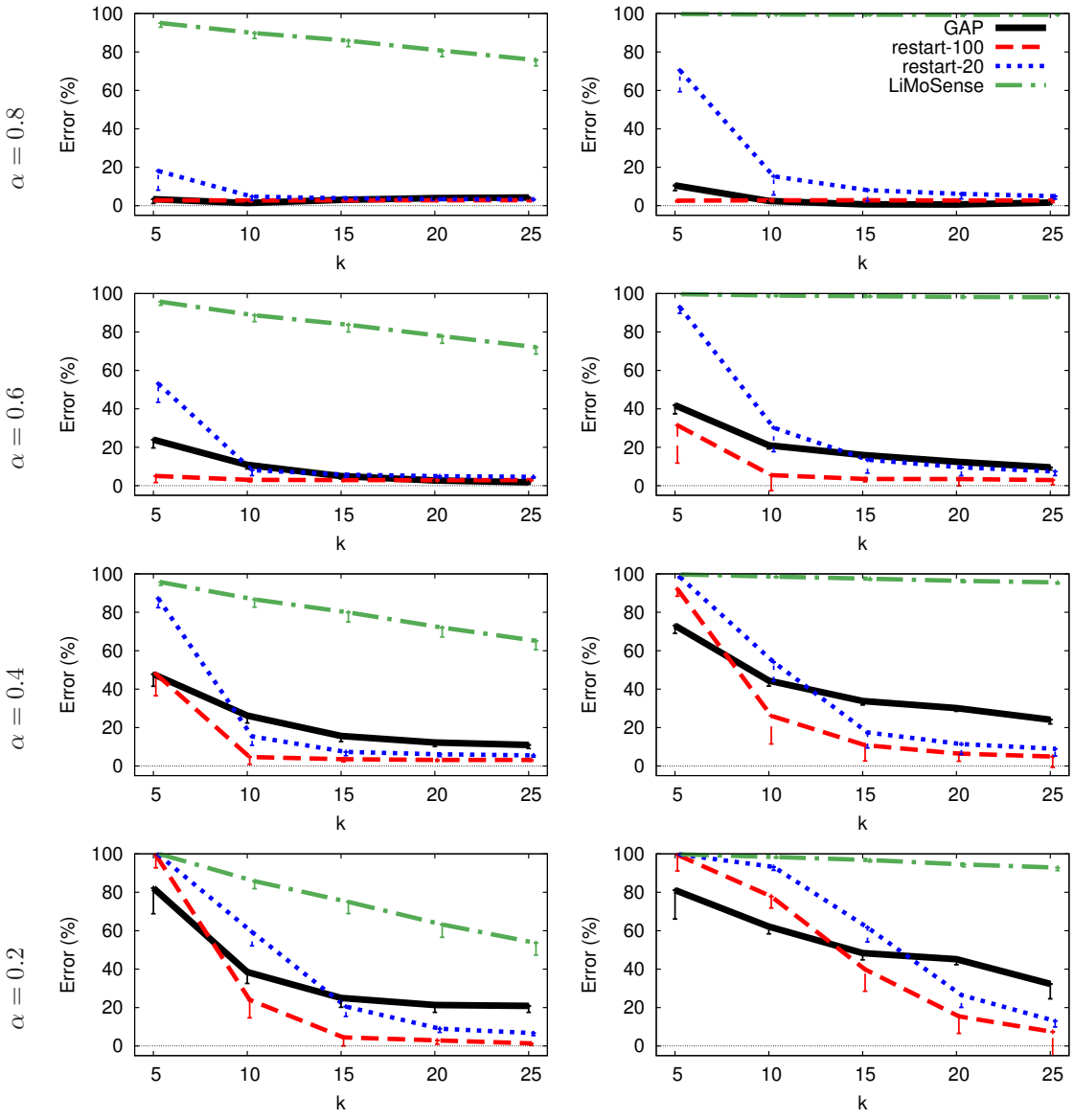

Figure 6: Error over the k-out topology with reliable root neighbors, N=10,000, fast churn (left column) and N=100,000, fast churn (right column), as a function of k. The proportion of online nodes is given by α. The label restart-m refers to restarted gossip with epoch length m.

Here, gossip protocols appear to be more sensitive to scale than GAP, unlike in the case of k-out networks.

5 Discussion and Conclusions

In this paper, we compared three different paradigms for global distributed aggregation: approaches based on a spanning tree, restarted gossip, and bookkeeping gossip. We argued that network size estimation is an appropriate problem for the purposes of this comparison. We stressed the role of different topologies, and shed light on the weak and strong points of the approaches.

Table 1 summarizes some of the conclusions we arrived at in the evaluation section. In our experiments the effective network size was constant, so the effect of the delay due to the epoch length remained hidden.

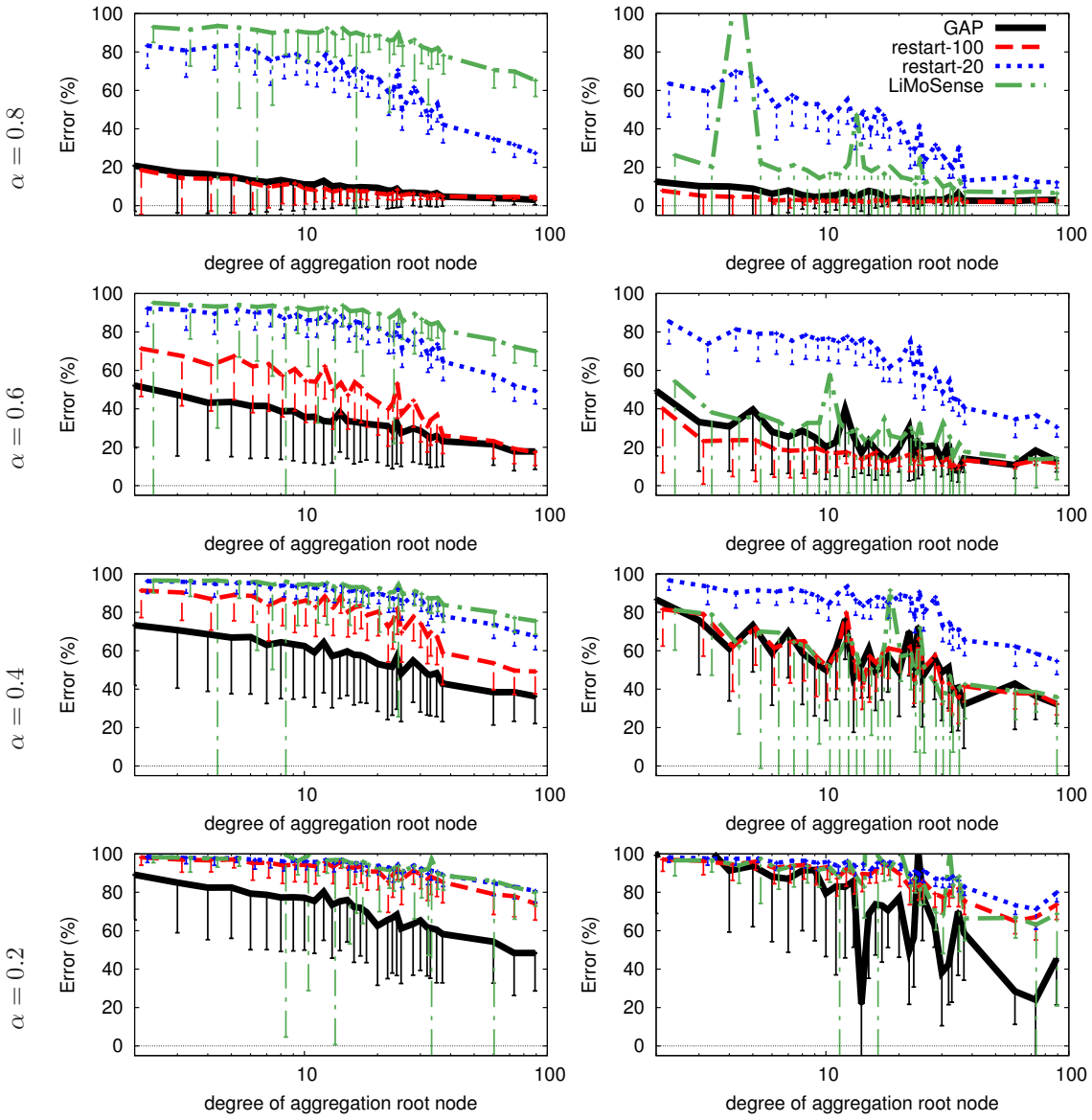

Figure 7: Error over the BA topology, fast churn (left column) and slow churn (right column), as a function of the degree of the node where the aggregation root is placed. The proportion of online nodes is given by α. The label restart-m refers to restarted gossip with epoch length m.

However, the epoch length must be chosen such that it lies in the range of the mixing time so as to allow for proper convergence. This means that restarted gossip will double the delay of bookkeeping gossip in

Table 1: Summary of conclusions.

	sensitivity to changing		**delay due to**	
	membership	**topology**	**convergence**	**epoch length**
spanning tree	moderate	high	diameter	none
bookkeeping gossip	high	high	mixing time	none
restarted gossip	moderate	none	mixing time	epoch length

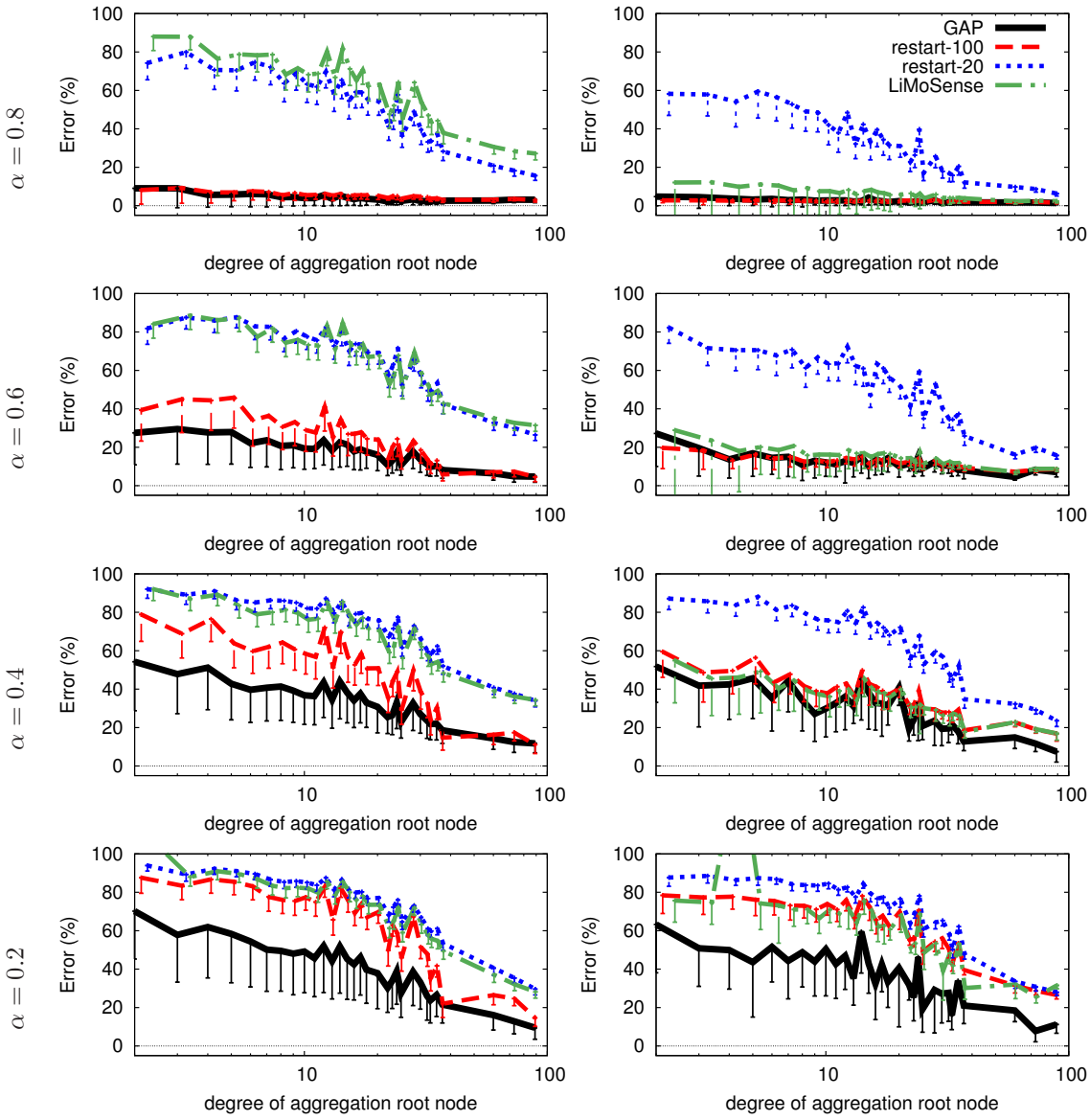

Figure 8: Error over the BA topology with reliable root neighbors, fast churn (left column) and slow churn (right column), as a function of the degree of the node where the aggregation root is placed. The proportion of online nodes is given by α. The label restart-m refers to restarted gossip with epoch length m.

the worst case, while it is not sensitive to a dynamic topology (due to not relying on failure detectors and neighborhood tables) and it is less sensitive to churn for the same reason.

As for the spanning tree, the convergence time of gossip (that depends on the mixing time) is typically at least an order of magnitude larger than the diameter in most topologies, even in the random k-out topology (which has a low mixing time), let alone more practical topologies. Our experiments clearly support this insight. This means that a spanning tree is much faster than the other methods, and its advantages mainly result from this property, along with the ability to self-repair equally quickly, when the topology is not too dynamic.

Nevertheless, we found that restarted gossip protocols consistently outperform the spanning tree on the random k-out network when k is sufficiently large and when the epoch length is long enough. The advantage of gossip is larger when churn is higher, but in that case gossip is even more sensitive to k. One

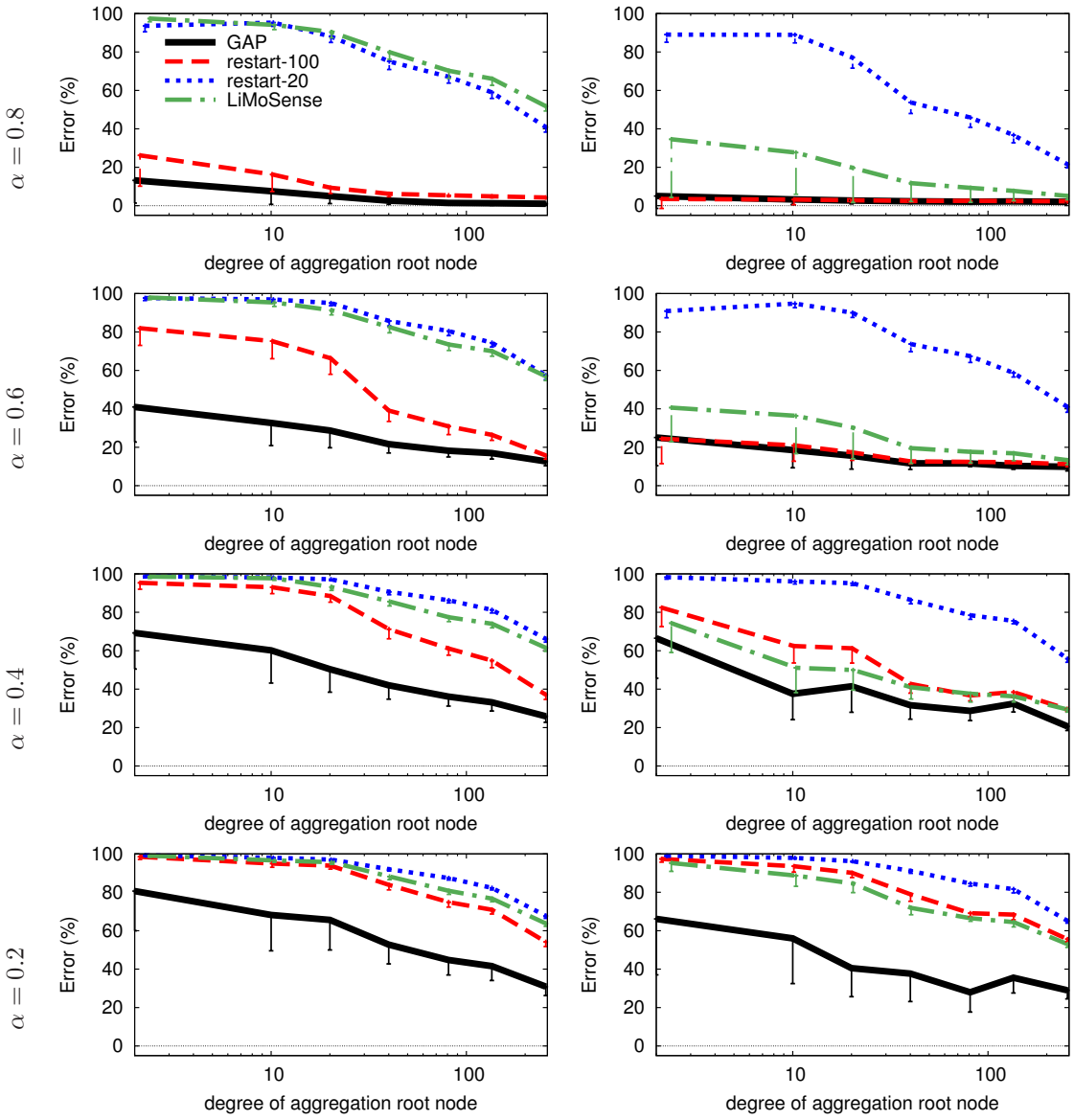

Figure 9: Error over the BA topology with reliable root neighbors, N=10,000, fast churn (left column) and slow churn (right column), as a function of the degree of the node where the aggregation root is placed. The proportion of online nodes is given by α. The label restart-m refers to restarted gossip with epoch length m.

reason for this behavior might be that although the spanning tree is faster in repairing its output, it is more sensitive to temporary failures, so in high continuous churn gossip is still preferable, but only when the mixing time of the network is optimal.

In the case of the scale free network the spanning tree is the clear winner. This is due to the larger mixing time that is associated with gossip protocols over such networks and also due to the fact that in such networks gossip is almost as sensitive to node failures as the spanning tree, since the network partitions much easier (i.e. when central nodes fail). Bookkeeping gossip is competitive with restarted gossip whenever churn is slow enough. However, bookkeeping gossip is not clearly preferable in any of the scenarios we examined in our study.

The size of the network also has a notable impact that is rather different depending on the type of the

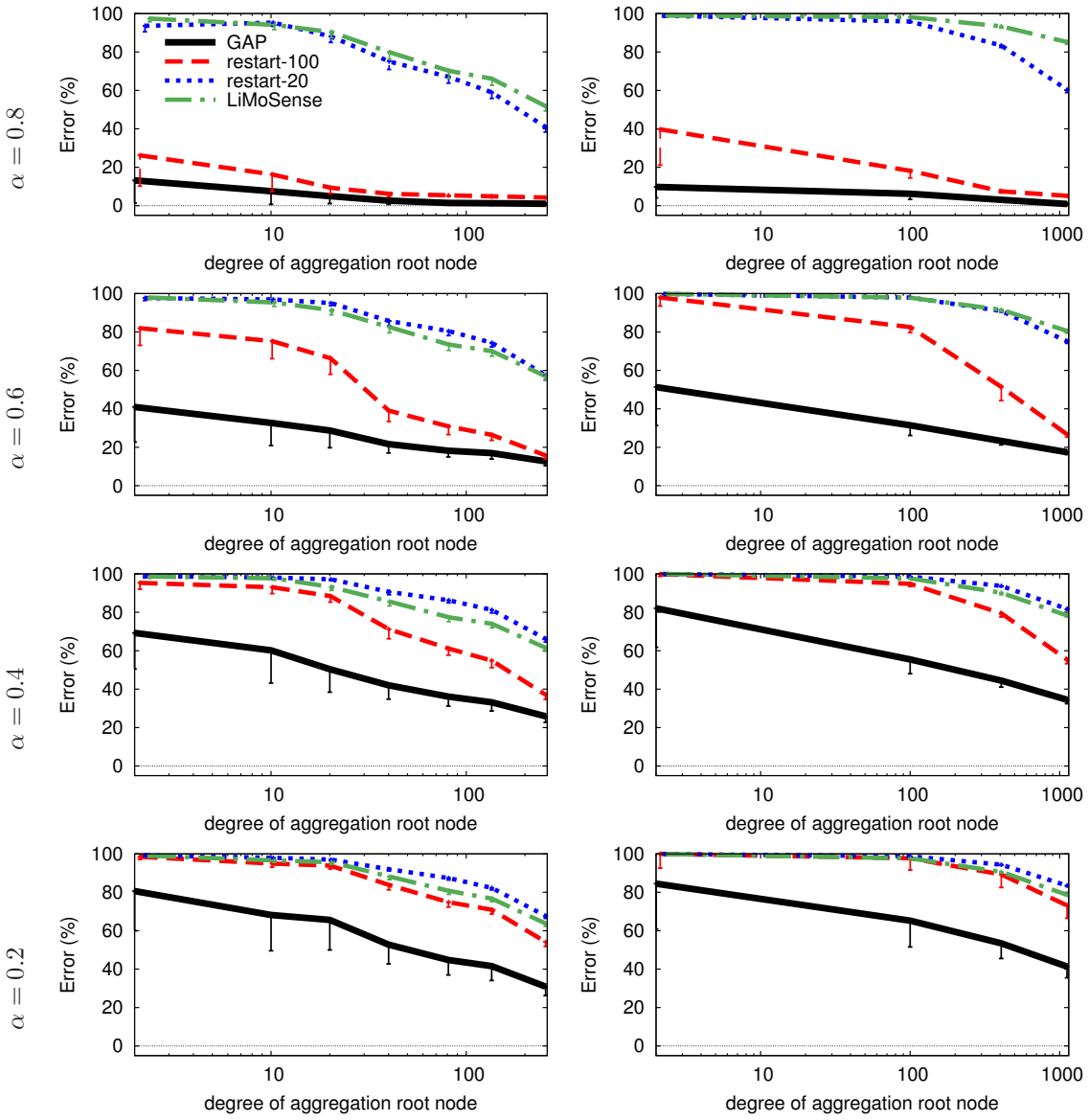

Figure 10: Error over the BA topology with reliable root neighbors, N=10,000, fast churn (left column) and N=100,000, fast churn (right column), as a function of the degree of the node where the aggregation root is placed. The proportion of online nodes is given by α. The label restart-m refers to restarted gossip with epoch length m.

network topology. In k-out networks spanning tree approaches suffer more from an increasing network size. This is because the number of corrections due to detecting a failed neighbor grows linearly with network size assuming a constant churn rate, so the root will experience an increasing amount of noise (although these effects average out to a large degree, see [11]). At the same time, gossip protocols depend mostly on mixing time that grows slowly with network size in the case of random networks. However, in the case of scale free networks or (even worse) trees, mixing time grows faster with network size. This makes the spanning tree a preferable choice even in large networks in topologies where the mixing time is large, since the increased convergence time of gossip protocols outweighs the the effect of the increased noise in the spanning tree.

Overall, when selecting the right protocol, one needs to consider the structure of the topology and the

patterns of dynamism in the membership (churn) and the topology itself. For dynamic topologies a restarted gossip protocol with the right epoch length is more suitable, and the spanning tree approach is quite clearly not suitable. If the topology is relatively stable, then the situation becomes very complex. With no failure, or with very low failure rates, the spanning tree is very clearly the best choice. What is more surprising is that a spanning tree approach is often preferable even in high churn, especially in topologies that have a high mixing time or that are more fragile to random failure. However, restarted gossip is preferable even in static topologies (with churn), if we can guarantee an optimal mixing time and robustness (for example, if we have a random overlay network).

If local sampling approximates the global aggregate well, we face a very different problem that requires a different approach for analysis. Nevertheless, results on global problems always represent a lower bound on performance. The ultimate solution is most likely a combination of gossip and tree approaches in an adaptive way, based on the automated detection of topology and dynamism properties. This is an interesting direction for future research.

References

[1] Jelasity M, Montresor A, Babaoglu O. Gossip-based aggregation in large dynamic networks. *ACM Transactions on Computer Systems* August 2005; **23**(3):219–252, doi:10.1145/1082469.1082470.

[2] Eyal I, Keidar I, Rom R. Limosense – live monitoring in dynamic sensor networks. *Algorithms for Sensor Systems*, *Lecture Notes in Computer Science*, vol. 7111, Erlebach T, Nikoletseas S, Orponen P (eds.). Springer Berlin / Heidelberg, 2012; 72–85, doi:10.1007/978-3-642-28209-6_7.

[3] Jesus P, Baquero C, Almeida P. Fault-tolerant aggregation by flow updating. *Distributed Applications and Interoperable Systems*, *Lecture Notes in Computer Science*, vol. 5523, Senivongse T, Oliveira R (eds.). Springer Berlin / Heidelberg, 2009; 73–86, doi:10.1007/978-3-642-02164-0_6.

[4] Mehyar M, Spanos D, Pongsajapan J, Low SH, Murray RM. Asynchronous distributed averaging on communication networks. *IEEE/ACM Trans. Netw.* 2007; **15**(3):512–520, doi:10.1109/TNET.2007.893226.

[5] Wuhib F, Dam M, Stadler R, Clemm A. Robust monitoring of network-wide aggregates through gossiping. *Proc. 10th IFIP/IEEE International Symposium on Integrated Management (IM 2007)*, Munich, Germany, 2007; 21–25, doi:10.1109/INM.2007.374787.

[6] Madden S, Franklin MJ, Hellerstein JM, Hong W. TAG: a tiny aggregation service for ad-hoc sensor networks. *Proceedings of the 5th Symposium on Operating Systems Design and Implementation (OSDI'02)*, Boston, Massachusetts, USA, 2002; 131–146.

[7] Gupta I, van Renesse R, Birman KP. Scalable fault-tolerant aggregation in large process groups. *Proceedings of the International Conference on Dependable Systems and Networks (DSN'01)*, IEEE Computer Society Press: Göteborg, Sweden, 2001; 433–442, doi:10.1109/DSN.2001.941427.

[8] Birman KP, van Renesse R, Vogels W. Scalable data fusion using astrolabe. *Proceedings of the Fifth International Conference on Information Fusion (FUSION 2002)*, vol. 2, 2002; 1434–1441, doi:10.1109/ICIF.2002.1020984.

[9] Dam M, Stadler R. A generic protocol for network state aggregation. *Proceedings of Radiovetenskap och Kommunikation (RVK'05)*, Linköping, Sweden, 2005.

[10] Prieto AG, Stadler R. A-gap: An adaptive protocol for continuous network monitoring with accuracy objectives. *IEEE Trans. on Netw. and Serv. Manag.* June 2007; **4**(1):2–12, doi:10.1109/TNSM.2007.030101.

[11] Krishnamurthy S, Ardelius J, Aurell E, Dam M, Stadler R, Wuhib FZ. Brief announcement: the accuracy of tree-based counting in dynamic networks. *ACM Symposium on Principles of Distributed Computing (PODC)*, Richa AW, Guerraoui R (eds.), ACM, 2010; 291–292, doi:10.1145/1835698.1835770.

[12] Jain N, Mahajan P, Kit D, Yalagandula P, Dahlin M, Zhang Y. Network imprecision: A new consistency metric for scalable monitoring. *Proceedings of the 8th USENIX Conference on Operating Systems Design and Implementation (OSDI'08)*, USENIX Association: Berkeley, CA, USA, 2008; 87–102.

[13] Le Merrer E, Kermarrec AM, Massoulie L. Peer to peer size estimation in large and dynamic networks: A comparative study. *Proceedings of the 15th IEEE International Symposium on High Performance Distributed Computing (HPDC'06)*, 2006; 7–17, doi:10.1109/HPDC.2006.1652131.

[14] Chitnis L, Dobra A, Ranka S. Aggregation methods for large-scale sensor networks. *ACM Trans. Sen. Netw.* April 2008; **4**(2):9:1–9:36, doi:10.1145/1340771.1340775.

[15] Nyers L, Jelasity M. Spanning tree or gossip for aggregation: A comparative study. *Euro-Par 2014*, *Lecture Notes in Computer Science*, vol. 8632, Silva F, Dutra I, Santos Costa V (eds.), Springer International Publishing, 2014; 379–390, doi:10.1007/978-3-319-09873-9_32.

[16] Dolev S, Israeli A, Moran S. Self-stabilization of dynamic systems assuming only read/write atomicity. *Distributed Computing* 1993; **7**(1):3–16, doi:10.1007/BF02278851.

[17] Kempe D, Dobra A, Gehrke J. Gossip-based computation of aggregate information. *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science (FOCS'03)*, IEEE Computer Society, 2003; 482–491, doi:10.1109/SFCS.2003.1238221.

[18] Montresor A, Jelasity M. Peersim: A scalable P2P simulator. *Proceedings of the 9th IEEE International Conference on Peer-to-Peer Computing (P2P 2009)*, IEEE: Seattle, Washington, USA, 2009; 99–100, doi:10.1109/P2P.2009.5284506. Extended abstract.

[19] Jelasity M, Voulgaris S, Guerraoui R, Kermarrec AM, van Steen M. Gossip-based peer sampling. *ACM Transactions on Computer Systems* August 2007; **25**(3):8, doi:10.1145/1275517.1275520.

[20] Roverso R, Dowling J, Jelasity M. Through the wormhole: Low cost, fresh peer sampling for the internet. *Proceedings of the 13th IEEE International Conference on Peer-to-Peer Computing (P2P 2013)*, IEEE, 2013, doi:10.1109/P2P.2013.6688707.

[21] Albert R, Barabási AL. Statistical mechanics of complex networks. *Reviews of Modern Physics* January 2002; **74**(1):47–97.

[22] Roozenburg J. Secure decentralized swarm discovery in Tribler. Master's Thesis, Parallel and Distributed Systems Group, Delft University of Technology 2006. URL `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.143.5153&rep=rep1&type=pdf`.

[23] Stutzbach D, Rejaie R. Understanding churn in peer-to-peer networks. *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement (IMC'06)*, ACM: New York, NY, USA, 2006; 189–202, doi:10.1145/1177080.1177105.

[24] Boyd S, Ghosh A, Prabhakar B, Shah D. Randomized gossip algorithms. *IEEE Transactions on Information Theory* 2006; **52**(6):2508–2530, doi:10.1109/TIT.2006.874516.

[25] Levin DA, Peres Y, Wilmer EL. *Markov Chains and Mixing Times*. American Mathematical Society, 2008.