Appendix A

Theorem 2

Given \( n \) peers \( p_1, \ldots, p_n \), let a signature \( s \) be published by some \( m \) peers with frequencies \( f_1, \ldots, f_m \), where \( m \leq n \). With at least probability \( 1 - \delta \), there is a round \( t_0 = O(\log(n) + \log(\frac{1}{\delta}) + \log(\frac{1}{\varepsilon})) \), such that in all rounds \( t \geq t_0 \), at peer \( p_i \), the relative error of the estimate of the average frequency of \( s \), i.e., \( \frac{1}{m} \sum_{i=1}^{m} f_i \), is at most \( \varepsilon \).

Proof. In this proof, we show that VanillaXGossip does not violate “mass conservation” and therefore the proof for Push-Sum holds for VanillaXGossip. Without loss of generality, suppose only one signature \( s \) exists in the network and a peer \( p \) is observed. Then \( p \) has either published \( s \) or has not.

Case 1: Suppose \( p \) has not published \( s \). Suppose \( p \) knows this fact, and starts with a sum and weight \( (0, 1) \). In round \( t \), \( p \) receives \( (f_s, w) \) where \( f_s = 0 \) such that in all the previous rounds \( p \) has only received messages from those peers that have not published \( s \). Suppose \( (0, w) \) be the sum and weight in round \( t - 1 \) at \( p \). Then \( p \) will compute its new sum to be \( f_s + 0 \) and weight \( w + w \) and sends \( \frac{f_s}{w} \) and \( \frac{w}{w} \) to another peer. Now VanillaXGossip resembles Push-Sum and mass conservation is preserved and the proof of Push-Sum holds.

But if \( p \) does not know the fact that \( s \) exists, and uses the placeholder \( \bot \) and starts with \( (0, 1) \) as the sum and weight for this placeholder signature. Suppose we replay the actions up to round \( t \). Now in round \( t \), \( p \) receives \( (f_s, w) \) where \( f_s > 0 \). Now the sum and weight based on \( \bot \) will be \( (0, w) \) in round \( t - 1 \) at \( p \). Peer \( p \) will compute its new sum to be \( f_s + 0 \) and weight \( w + w \) and sends \( \frac{f_s}{w} \) and \( \frac{w}{w} \) to another peer. So even when \( p \) does not know about the existence of \( s \), it can arrive at the right sum and weight in round \( t \) to guarantee mass conservation.

Case 2: Suppose \( p \) has published \( s \). Without loss of generality, suppose in round \( t \), \( p \) receives the placeholder signature with \( (0, w) \) from some peer \( q \). This means that so far \( q \) has received messages from peers that do not know about \( s \) to begin with. Then \( p \) computes the sum and weight as \( f_s + 0 \) and \( w + w \). This would be the same if the peers that have sent a message to \( q \) (including \( q \)) until round \( t - 1 \), started with \( (0, 1) \) for signature \( s \) if they assumed that \( s \) existed in the network. Then mass conservation is guaranteed and the proof of Push-Sum holds.

From cases 1 and 2, we can conclude that for any \( s \), mass conservation holds in VanillaXGossip and therefore, the proof of Push-Sum holds.

Theorem 3

Given \( n \) peers \( p_1, \ldots, p_n \), in a network, let a signature \( s \) be published by some \( m \) peers with frequencies \( f_1, \ldots, f_m \), where \( m \leq n \). Suppose \( p_i \) belongs to a team that gossips \( s \) after applying LSH on \( s \). Let \( \Delta \) denote the team size. With at least probability \( 1 - \delta \), there is a round \( t_0 = O(\log(\Delta) + \log(\frac{1}{\delta}) + \log(\frac{1}{\varepsilon})) \), such that in all rounds \( t \geq t_0 \), at peer \( p_i \), the relative error of the estimate of the average frequency of \( s \), i.e., \( \frac{1}{m} \sum_{i=1}^{m} f_i \), is at most \( \varepsilon \).

Proof. Suppose \( h_s = (h_{s1}, \ldots, h_{sk}) \) denotes the output of LSH on \( s \). Without loss of generality, consider the team \( h_{si} \). During initialization, any peer that published an XML document whose signature is \( s \), will send \( (s, (f, w)) \) to a member of \( h_{si} \). At the end of the initialization phase, mass conservation holds for \( s \) w.r.t. team \( h_{si} \). This is because the average of the frequency of \( s \) across all the members of team \( h_{si} \) is the true average, and the sum of weights for \( s \) is \( \Delta \). During the execution phase, \( s \) is gossipied by the members of team \( h_{si} \) and mass conservation is preserved like in VanillaXGossip due to the use of the special multiset \( \perp_{h_{si}} \). Now the situation is identical to VanillaXGossip except that the number of peers involved in computing the average is \( \Delta \). Hence the above theorem holds.

Theorem 4

Given an XPath query \( q \), VanillaXGossip can estimate the cardinality of \( q \) with a relative error of at most \( \varepsilon \) and a probability of at least \( 1 - \delta \) in \( O(\log(n) + \log(\frac{1}{\varepsilon}) + \log(\frac{1}{\delta})) \) rounds.

Proof. From Theorem 2, we know that the frequency of a signature in \( R \) can be estimated with a relative error of at most \( \varepsilon \) and confidence \( (1 - \delta) \) in \( O(\log(n) + \log(\frac{1}{\varepsilon}) + \log(\frac{1}{\delta})) \), where \( n \) denotes the number of peers in the network. In VanillaXGossip, the frequencies of \( q \) signatures in \( R \) is used to compute the cardinality estimate of \( q \), and therefore, the total relative error is at most \( r \).

Theorem 5

Given an XPath query \( q \), suppose \( q_{min} \) denotes the minimum similarity between \( q \)’s signature and a signature in \( R \). XGossip can estimate the cardinality of \( q \) with a relative error of at most \( \varepsilon \) and a probability of at least \( \alpha \cdot (1 - \delta) \) in \( O(\log(\Delta) + \log(\frac{1}{\varepsilon}) + \log(\frac{1}{\delta})) \) rounds, where \( \alpha = 1 - (1 - q_{min})^{k \cdot l} \), and \( k \) and \( l \) denote the parameters of LSH.

Proof. From Theorem 2, we know that the frequency of a signature in \( R \) can be estimated with a relative error of at most \( \varepsilon \) and confidence \( (1 - \delta) \) in \( O(\log(\Delta) + \log(\frac{1}{\varepsilon}) + \log(\frac{1}{\delta})) \), where \( \Delta \) denotes the team size. In XGossip, the frequencies of \( r \) signatures in \( R \) is used to compute the cardinality estimate of \( q \), and therefore, the total relative error is at most \( r \).

The confidence of the estimate also depends on the properties of LSH. Because \( q_{min} \) denotes the minimum similarity between \( q \)’s signature and a signature in \( R \), the probability of finding all signatures in \( R \) by contacting \( k \) teams at query time is at least \( \alpha = 1 - (1 - q_{min})^{k \cdot l} \), where \( k \) and \( l \) are the parameters of LSH. (Suppose a proxy signature is used such that the minimum similarity between it and a signature in \( R \) is \( q_{min} \). Then \( \alpha = 1 - (1 - p_{min})^{k \cdot l} \).) Therefore, the net confidence is at least \( \alpha \cdot (1 - \delta) \).

Corollary 1

XGossip can estimate the cardinality of \( q \) with a relative error of at most \( \varepsilon \) and a probability of at least \( (1 - \delta) \) in \( O(\log(\Delta) + \log(\frac{1}{\varepsilon}) + \log(\frac{1}{\delta})) \) rounds.

Proof. To achieve the same confidence as VanillaXGossip, i.e., \( (1 - \delta) \), the following equations must hold:

\[
(1 - \delta) = \alpha \cdot (1 - \delta)^{k \cdot l}
\]

\[
\alpha = \frac{\alpha + \delta - 1}{\alpha}
\]

Therefore, XGossip achieves a relative error of at most \( \varepsilon \) with a probability of at least \( (1 - \delta) \) in \( O(\log(\Delta) + \log(\frac{1}{\varepsilon}) + \log(\frac{1}{\delta})) \) rounds.
Appendix B

Accuracy of Cardinality Estimation

Figure 26 shows the accuracy obtained by XGossip after 10 rounds for 500, 1000, and 2000 peers.