Energy Cost Model for Frequent Item Set Discovery in Unstructured P2P Networks *

Emrah Cem[1], Ender Demirkaya[2],
Ertem Esiner[1], Burak Ozaydin[1], Oznur Ozkasap[1]

[1] Department of Computer Engineering, Koc University, Istanbul, Turkey
[2] Department of Computer Engineering, Bilkent University, Ankara, Turkey

Abstract. For large scale distributed systems, designing energy efficient protocols and services has become as significant as considering conventional performance criteria like scalability, reliability, fault-tolerance and security. We consider frequent item set discovery problem in this context. Although it has attracted attention due to its extensive applicability in diverse areas, there is no prior work on energy cost model for such distributed protocols. In this paper, we develop an energy cost model for frequent item set discovery in unstructured P2P networks. To the best of our knowledge, this is the first study that proposes an energy cost model for a generic peer using gossip-based communication. As a case study protocol, we use our gossip-based approach ProFID for frequent item set discovery. After developing the energy cost model, we examine the effect of protocol parameters on energy consumption using our simulation model on PeerSim and compare push-pull method of ProFID with the well-known push-based gossiping approach. Based on the analysis results, we reformulate the upper bound for the peer's energy cost.

Keywords: energy cost model, energy efficiency, peer-to-peer, gossip-based, epidemic, frequent items.

1 Introduction

Frequent items in a distributed environment can be defined as items with global frequency above a threshold value, where global frequency of an item refers to the sum of its local values on all peers. Frequent Item Set Discovery (FID) problem has attracted significant attention due its extensive applicability in diverse areas such as P2P networks, database applications, data streams, wireless sensor networks, and security applications.

In this study, we propose and develop an energy cost model for a generic peer using gossip-based communication for FID. Gossip-based or epidemic mechanisms are preferred in several distributed protocols [7, 6] for their ease of deployment, simplicity, robustness against failures, and limited resource usage. In

* This work was partially supported by the COST (European Cooperation in Science and Technology) framework, under Action IC0804, and by TUBITAK (The Scientific and Technical Research Council of Turkey) under Grant 109M761.

terms of their power usage, the efficiency of three models of epidemic protocols, namely basic epidemics, neighborhood epidemics and hierarchical epidemics, has been examined in [8]. Basic epidemics that requires full membership knowledge of peers was found to be inefficient in its power usage. It has been shown that; in neighborhood epidemics, peer's power consumption amount is independent of population size. For hierarchical epidemics, power usage increases with population size. In fact, [8] is the only study that considers power awareness features of epidemic protocols. However, it evaluates different epidemics through simulations only and provides results on latency and power (proportional to the gossip rate). Moreover, effects of gossip parameters such as fan-out and maximum gossip message size were not investigated. In contrast, our study is the first one that proposes an energy cost model for a generic peer using gossip-based communication like in ProFID protocol, and examines the effect of protocol parameters to characterize energy consumption. As a case study protocol, we use our gossip-based approach ProFID for frequent item set discovery [4]. It uses a novel atomic pairwise averaging for computing average global frequencies of items and network size, and employs a convergence rule and threshold mechanism. Due to the page limitation, we refer interested reader to [4] for details of the protocol.

This paper is organized as follows. Section 2 develops energy cost model for a gossip-based peer used in our protocol. Section 3 analyzes the effect of protocol parameters, compares push-pull method of ProFID with the well-known push-based gossiping that we adapted to frequent item set discovery, and reformulates the peer's energy cost. Finally, section 4 states conclusions and future directions.

2 Energy Cost Model

ProFID protocol depends on three main components of operations performed by each peer: energy consumed while (1) computing new state, (2) sending messages and (3) receiving messages. Inspired by studies [9, 3], we propose an energy cost model for a generic peer using gossip-based communication in ProFID. In study [9], energy cost models for client-server and publish-subscribe styles were developed. Then, application and platform specific model parameters were also taken into consideration and energy prediction model was developed. Work of [3] introduces a quorum-based model to compute energy costs of read and write operations in replication protocols, and proposes an approach to reduce the energy cost of tree replication protocol. Different than these prior works, we develop energy cost model for a peer using gossip-based communication and consider the effects of gossip parameters on the cost representation.

We start with the analysis of the energy consumption during an atomic pairwise averaging operation between peers P_i and P_j Different operations consuming energy are explained in Table 1. During an atomic pairwise averaging, energy cost of a peer that initiates a gossip (*gossip starter*) is represented by:

$$E_{gossipStarter} = E_{send} + E_{receive} + E_{compStarter} \quad (1)$$

Table 1: Different operations that consume energy

Value	Description
E_{send}	Energy required to send the item tuple
E_{recv}	Energy required to receive the item tuple
$E_{compStarter}$	Energy required to choose tuple to send and update the state
$E_{compTarget}$	Energy required to compute the average and prepare the tuple to send

On the other hand, energy cost of the gossip target can be formulated as follows:

$$E_{gossipTarget} = E_{receive} + E_{send} + E_{compTarget} \tag{2}$$

Note that $E_{compTarget}$ and $E_{compStarter}$ are both proportional to the gossip message size, and they can simply be represented as E_{comp}. Hence, $E_{i,j}$ (the energy consumption of a peer P_i during an atomic pairwise averaging with P_j) can be written as:

$$E_{i,j} = E_{send,j} + E_{receive,j} + E_{comp} + C \tag{3}$$

where $E_{send,j}$ is the energy consumed while sending a gossip message to P_j, $E_{receive,j}$ is the energy consumed while receiving a gossip message from P_j, and E_{comp} is the local computation of the peer. Note that this is the energy cost of a peer that performs an atomic pairwise averaging operation. In real network scenarios, energy consumption may include extra factors such as CPU's energy consumption during I/O. Hence, a constant C is added to the equation.

To represent the energy cost of a gossip-based peer during an atomic pairwise averaging operation, the formula was given with respect to the basic conditions (gossip to one neighbor, one round, one item). Step by step, we now extend this cost model of a peer for the ProFID protocol. A peer may initiate multiple gossip operations during a single round depending on the *fanout* value as well as it may become gossip target multiple times. The energy cost of P_i that gossips a single item tuple in a round can be formulated as:

$$E_{P_i}(\text{single round, single item}) = \sum_{j \in V \cup W} E_{i,j} \tag{4}$$

where V is the set of neighbors chosen by P_i as gossip targets, and W is the set of neighbors that initiates an atomic pairwise averaging with P_i. Note that the number of elements in V corresponds to the *fanout* value.
In general, a gossip message comprises multiple item tuples whose number is upper-bounded by *maximum message size* (*mms*) parameter. Since $E_{send,j}$ and $E_{receive,j}$ are the energies consumed while sending and receiving a single tuple respectively, total energy consumed during a gossip round would linearly increase with the *mms*. Hence, energy cost of P_i in a round can be expressed as:

$$E_{P_i}(\text{single round}) \leq mms \cdot \sum_{j \in V \cup W} E_{i,j} \tag{5}$$

Since a peer repeats those operations in every round, number of rounds R would increase the energy cost of a peer proportionally. Hence, the overall energy cost of P_i can be written as:

$$E_{P_i} \leq R \cdot mms \cdot \sum_{j \in V \cup W} E_{i,j} \tag{6}$$

3 Analysis and Results

We have developed a simulation model for ProFID protocol [2] on PeerSim simulator [1] and analyzed the effects of protocol parameters on the energy consumption. As presented in Eq. 6, energy cost of a peer is proportional to the convergence time, that is the number of rounds R. In this section, we analyze the effects of protocol parameters on R, compare push-pull based method of ProFID with the well-known push-based gossiping, evaluate the effects of convergence parameters on frequency error (i.e. the percentage of items which were identified as frequent though they are actually not) and reformulate the upper bound of the overall energy cost of a peer in terms of protocol parameters.

We performed our evaluations through extensive large-scale distributed scenarios (up to 30,000 peers) on PeerSim. We tested different topologies such as random topology and scale-free Barabasi-Albert topology with average degree 10. All the data points presented in graphs are the average of 50 experiments. The default values of parameters used in the experiments are given in Table 2.

Table 2: Default parameter values

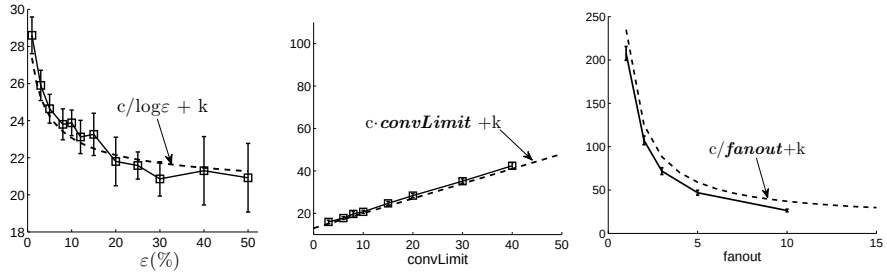
Parameter	Value	Parameter	Value	Parameter	Value
N	1000	M(number of items)	100	*convLimit*	10
ε	10	*mms*	100	*fanout*	1

Convergence Parameters (convLimit, ε): Convergence parameters are used for self-termination of peers and they have direct effects on R. Fig. 1a shows that R is inversely proportional to $log\varepsilon$. This is because *convCounter* will be incremented with less chance and it will take longer time to reach *convLimit*. However, R is directly proportional to *convLimit* as depicted in Fig. 1b, and this is because *convCounter* needs to be incremented more to take convergence decision.

Fanout: Intuitively, increasing *fanout* will cause to consume more energy in a single round. On the other hand, algorithm will converge faster since a peer exchanges its state with more peers in a single round. Fig. 1c depicts that *fanout* has an inverse proportion with R. Note also that *fanout* has a direct proportion with the upper bound given in Eq. 6 since *fanout* is the cardinality of set V.

Gossip message size: Parameter *mms* is the upper bound for a gossip message size in terms of number of <item,frequency> tuples. Large *mms* means more state information is sent in a single gossip message. On one hand, this causes faster convergence, but on the other hand, the energy consumption of sending a single gossip message increases. Results in Fig. 2a verify that *mms* is inversely

proportional to R. Note also that *mms* is directly related with the energy cost of a peer in a single round, and these cancel each other in our cost formulation. Recall that ProFID assumes each peer knows about its neighboring peers only and gossips with them, and hence it is based on neighborhood epidemics. In this respect, our results are also consistent with [8] that reports the efficiency of neighborhood epidemics in its power usage.



(a) (b) (c)

Fig. 1: Effects of (a) ε on R (b) *convLimit* on R (c) *fanout* on R

Comparison with Adaptive Push-sum: We have compared ProFID with the Push-sum approach [5] to observe different gossip-based approaches as a solution to the FID problem. In order to compute aggregates of items, Push-sum protocol assumes that all peers are aware of all items in the network which is not practical for the case of FID. For this reason, we have developed an Adaptive Push-sum protocol on PeerSim by modifying the Push-sum algorithm and included the convergence rule in order to adapt it to the FID problem. As depicted in Fig. 2b, ProFID converges faster than Adaptive Push-Sum algorithm in all different fanout values. We also observed that ProFID outperforms Adaptive Push-Sum in terms of message complexity in these simulations.

Energy Cost and Frequency Error in Terms of Protocol Parameters: Combining the experimental analysis results, effects of protocol parameters on convergence time R can be represented as:

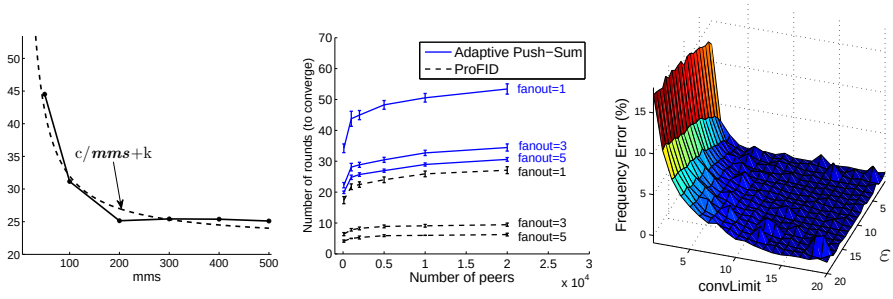
$$R \approx (1/log\varepsilon) \cdot logN \cdot convLimit \cdot (1/fanout) \cdot (1/mms) \tag{7}$$

Based on these findings above, we can reformulate the energy cost of P_i (in Eq. 6) as follows:

$$E_{P_i} \leq (1/log\varepsilon) \cdot logN \cdot convLimit \cdot (1/fanout) \cdot (\sum_{j \in V \cup W} E_{i,j}) \tag{8}$$

We should also consider the frequency error while minimizing the energy cost since obtaining unreasonable results with low energy cost would not be meaningful. Frequency error can be written in terms of protocol parameters by combining experimental result shown in Fig. 2c as follows:

$$FrequencyError \approx \varepsilon / convLimit \tag{9}$$



(a) (b) (c)

Fig. 2: (a) Effect of *mms* on frequency error (b) ProFID versus Adaptive Push-Sum (c) Effect of convergence parameters on frequency error

4 Conclusions and Future Work

Frequent item set discovery problem in P2P networks is relevant for several distributed services such as cache management, data replication, sensor networks and security. Our study is the first one that introduces and develops an energy cost model for a generic peer using gossip-based communication. Different than the prior works, we also studied the effect of protocol parameters through extensive large-scale simulations, compared push-pull and push-based gossiping methods. As future work, we plan to deploy our protocol on PlanetLab and analyze its energy cost on this network testbed. We also aim to extend our energy cost model to hierarchical gossip approaches.

References

1. The Peersim simulator, `http://peersim.sf.net`
2. The ProFID implementation, https://sites.google.com/a/ku.edu.tr/emrah-cem/projects/profid
3. Basmadjian, R., de Meer, H.: An approach to reduce the energy cost of the arbitrary tree replication protocol. In: Proc. of e-Energy. pp. 151–158 (2010)
4. Cem, E., Ozkasap, O.: Profid: Practical frequent item set discovery in peer-to-peer networks. In: Proc. of ISCIS. pp. 199–202 (2010)
5. Kempe, D., Dobra, A., Gehrke, J.: Gossip-based computation of aggregate information. In: Proc. of FOCS. pp. 482–491 (2003)
6. Ozkasap, O., Caglar, M., Yazici, E., Kucukcifci, S.: An analytical framework for self-organizing peer-to-peer anti-entropy algorithms. Performance Evaluation Journal, Elsevier Science 67(3), 141–159 (2010)
7. Ozkasap, O., Genc, Z., Atsan, E.: Epidemic-based reliable and adaptive multicast for mobile ad hoc networks. Computer Networks 53 (2009)
8. van Renesse, R.: Power-aware epidemics. In: Proc. of IEEE Symposium on Reliable Distributed Systems (2002)
9. Seo, C., Edwards, G., et al.: A framework for estimating the energy consumption induced by a distributed system's architectural style. In: Proc. of SAVCBS (2009)