Refined Mean Field Analysis: The Gossip Shuffle Protocol Revisited

Nicolas Gast, Diego Latella, Mieke Massink

Refined Mean Field Analysis: The Gossip Shuffle Protocol Revisited*

Nicolas Gast[1], Diego Latella[2], and Mieke Massink[2]

[1] INRIA, University Grenoble Alpes, France
[2] Consiglio Nazionale delle Ricerche - Istituto di Scienza e Tecnologie dell'Informazione 'A. Faedo', CNR, Italy

Abstract. Gossip protocols form the basis of many smart collective adaptive systems. They are a class of fully decentralised, simple but robust protocols for the distribution of information throughout large scale networks with hundreds or thousands of nodes. Mean field analysis methods have made it possible to approximate and analyse performance aspects of such large scale protocols in an efficient way that is *independent* of the number of nodes in the network. Taking the gossip shuffle protocol as a benchmark, we evaluate a recently developed *refined* mean field approach. We illustrate the gain in accuracy this can provide for the analysis of medium size models analysing two key performance measures: replication and coverage. We also show that refined mean field analysis requires special attention to correctly capture the coordination aspects of the gossip shuffle protocol.

Keywords: Mean Field; Collective Adaptive Systems; Discrete Time Markov Chains; Gossip protocols; Self-organisation.

1 Introduction and Related Work

Many collective adaptive systems rely on the decentralised distribution of information. Gossip protocols have been proposed as a paradigm that can provide a stable, scalable and reliable method for such decentralised spreading of information [22,6,3,16,9,8,4,2,21]. The basic mechanism of information spreading followed by a gossip shuffle protocol is that nodes exchange part of the data they keep in their cache with randomly selected peers in pairwise synchronous communications on a regular basis.

Interesting performance aspects of such gossip protocols are the replication of a newly inserted fresh data element in a network and the dynamics of network coverage. Replication of a data element occurs when nodes exchange the data element in pairwise communication. Network coverage concerns the the fraction of the population of network nodes that have "seen" the data element since its

* This research has been partially supported by the Italian MIUR project PRIN 2017FTXR7S "IT-MaTTerS" (Methods and Tools for Trustworthy Smart Systems).

introduction into the network, even if they may no longer have it in their cache due to further exchanges with other peers.

Traditionally, these performance measures have been studied based on simulation models. However, when large populations of nodes are involved, such simulations may be very resource consuming. Recently these protocols have been studied using classic mean field approximation techniques [2,1]. In that classic approach the full stochastic model of a gossip network, i.e. one in which each node is modelled individually, is replaced by a much simpler model in which the pairwise synchronous interactions between individual nodes are replaced by the average effect that all those interactions have on a single node and then the model of this single node is studied in the context of the overall average network behaviour. Of course, the average effects may change over time as nodes change their local states. This is taken into account in a mean field model by letting the probabilities of interactions depend on the fraction of nodes that are in a particular local state. Compared to traditional simulation methods, mean field approximation techniques scale very well to large populations because these techniques are *independent of the exact population size*[3] allowing analysis that is orders of magnitudes faster than discrete event simulation. This method of derivation of a mean field model from a large population of interacting objects relies on what is known as the assumption of "propagation of chaos" (also called "statistical independence" or "decoupling of joint probabilities") [19,7,10,17]. The assumption is based on the fact that when the number of interacting nodes becomes very large, their interactions tend to behave as if they were statistically independent.

In this paper we revisit an analysis of the gossip shuffle protocol by Bahkshi et al. in [4,2,1] by using a *refined* mean field approximation for *discrete time population models* that we developed in [12,13], and which was in turn inspired by an earlier result for continuous time population models presented in [11].

Contributions The main contribution of this short paper (full version in [14]) is a novel benchmark (clock-synchronous) DTMC population model of the gossip shuffle protocol analysed using refined mean field analysis [12,13]. In particular:

- We show that, by using the refined mean field, a more accurate approximation can be obtained, compared to classical mean field approximation, for *medium size* populations for this gossip protocol, but that this requires a *novel model* that reflects the synchronisation effects of the pairwise interaction of the original protocol.
- The refined mean field results we obtained are very close both to those of independent Java based simulation from the literature in [2] (taken as "ground truth" for comparison with our results) and to those of the event simulation of the model itself, but several *orders of magnitude faster* and *independent* of the system size.

[3] As long as this size is large enough to obtain a sufficiently accurate approximation. The computational complexity of these techniques *do* depend on the number of local states of an object in a population.

Like classic mean field approaches, the refined approach is also highly scalable and computationally non-intensive. Therefore it is an interesting candidate for being integrated with other analysis approaches such as (on-the-fly) mean field model checking [18], which is planned in future work. The current study aims at providing further insight in the feasibility of applying the refined mean field approach, that implies the use of symbolic differentiation, on larger benchmark examples and in the possible complications of such an analysis that need to be taken into consideration.

2 Benchmark Gossip Shuffle Protocol

We consider the gossip shuffle protocol described in [15,1,2]. This particular version has been extensively studied by Bahkshi et al., leading to an analytical model of the gossip protocol [3], a classical mean field model [2] and a Java implementation[4] of a simulator for the protocol [2,1], which makes it a very suitable candidate of a real-world application that allows for the comparison of our results with those available in the literature. Fig. 1 recalls the pseudo code of a generic shuffle protocol (adapted from [1]). Further details can be found in [2,1].

while true **do**
wait (Δt time units)
B := randomPeer()
s_A := itemsToSend(c_A);
send s_A to B;

s_B := receive($\cdot$);
c_A := itemKeep($c_A \backslash (s_A \backslash s_B), s_B \backslash c_A$);

(a) An active node A

while true **do**

s_A := receive($\cdot$);
s_B := itemsToSend(c_B);
send s_B to sender(s_A);
c_B := itemKeep($c_B \backslash (s_B \backslash s_A), s_A \backslash c_B$);

(b) A passive contacted node B

Fig. 1. Pseudo code of a generic shuffle protocol (adapted from [1]). c_A and s_A denote the cache and selection of active node A. Similarly, c_B and s_B denote those of passive node B. $\Delta t = G_{max}$. The operation 'itemsToSend(c_i)' selects the items to be sent from the cache c_i. The operation 'itemKeep(c,s)' in node A decides which items to keep in the cache (c) removing from the cache those selected for sending (s_A) except those that where received from B (s_B), and adding to those the elements from s_B that were not yet in the cache of A. Similarly for the operation in node B.

Two main key measures that are of interest for this protocol are the transient aspects of the *replication* of a newly introduced element in the network and that of the *coverage* of the network, i.e. the fraction of network nodes that have seen

[4] We thank Rena Bahkshi for sharing her Java simulator source code with us.

the new data element when time is passing. These measures depend on a number of characteristics of the network. In the following we use N to denote the size of the network, i.e. the number of gossiping nodes, n to denote the number of *different* data items in the network, c to denote the size of the cache and s to denote the size of the selected items from the cache to be exchanged with a neighbour. In the context of this work, and for comparison with the results presented in [1], the network is assumed to be *fully connected*. We consider a discrete time variant of the protocol with a maximal delay between two subsequent active data-exchanges of a node denoted by G_{max}.

3 Background

In the sequel we use theoretical results on discrete time mean field approximation [19,7,12]. We briefly recall the notation and main results in the following. We consider a population model of a system composed of $0 < N \in \mathbb{N}$ identical interacting objects, i.e. a (model of a) system of *size* N. We assume that the set $\{0, \ldots, n-1\}$ of local states of each object is finite; we refer to [12] for a discussion on how to deal with infinite dimensional models. Time is *discrete* and the behaviour of the system is characterised by a (time homogeneous) *discrete time Markov chain* (DTMC) $X^{(N)}(t) = (X_1^{(N)}(t), \ldots, X_N^{(N)}(t))$, where $X_i^{(N)}(t)$ is the state of object i at time t, for $i = 1, \ldots, N$.

The *occupancy measure vector* at time t of the model is the row-vector DTMC $M^{(N)}(t) = (M_0^{(N)}(t), \ldots, M_{n-1}^{(N)}(t))$ where, for $j = 0, \ldots, n-1$, the stochastic variable $M_j^{(N)}(t)$ denotes the *fraction* of objects in state j at time t, over the total population of N objects:

$$M_j^{(N)}(t) = \frac{1}{N} \sum_{i=1}^{N} 1_{\{X_i^{(N)}(t)=j\}}$$

and $1_{\{x=j\}}$ is equal to 1 if $x = j$ and 0 otherwise. At each time step $t \in \mathbb{N}$ each object performs a local transition, possibly changing its state. The transitions of any two objects are assumed to be independent from each other, while the transition probabilities of an object may depend also on $M(t)$, thus, for large N, the probabilistic behaviour of an object is characterised by the one-step transition probability $n \times n$ matrix $\mathbf{K}(m)$, where $\mathbf{K}_{ij}(m)$ is the probability for the object to jump from state i to state j when the occupancy measure vector is $m \in \mathcal{U}^n$, the unit simplex of $\mathbb{R}^n_{\geq 0}$, that is, $\mathcal{U}^n = \{m \in [0,1]^n \mid \sum_{i=1}^n m_i = 1\}$. In this paper, for simplicity, we assume $\mathbf{K}(m)$ to be a continuous function of m that does not depend on N. In the sequel, for reasons of presentation, we provide a graphical specification of the relevant models. The computation of matrix $\mathbf{K}(m)$ from such a model specification is straightforward.

3.1 Discrete Time Classical Mean Field Approximation

Below we recall Theorem 4.1 of [19] on classic mean field approximation, under the simplifying assumptions mentioned above:

Theorem 4.1 of [19] (Convergence to Mean Field) *Assume that the initial occupancy measure $M^{(N)}(0)$ converges almost surely to the deterministic limit $\mu(0)$. Define $\mu(t)$ iteratively by (for $t \geq 0$):*

$$\mu(t+1) = \mu(t)\,\mathbf{K}(\mu(t)). \tag{1}$$

Then for any fixed time t, almost surely, $\lim_{N\to\infty} M^{(N)}(t) = \mu(t)$.

The above result thus allows one to use, for *large* N, a *deterministic* approximation μ of the average behaviour of a discrete population model.

3.2 Discrete Time Refined Mean Field Approximation

The following corollary illustrates the relationship between the refined mean field result and the classic convergence theorem:

Corollary 1(i) of [12] *Under the assumptions of* **Theorem 1 of [12]**, *it holds that for any coordinate i and any time-step $t \in \mathbb{N}$*

$$\mathbb{E}\left[M_i^{(N)}(t)\right] = \mu_i(t) + \frac{V_i(t)}{N} + o\left(\frac{1}{N}\right).$$

In other words, the expected value of the fraction of the objects in local state i of the full stochastic model with population size N at time t, is equal to the classic limit mean field value $\mu_i(t)$ plus a factor $V_i(t)$, divided by the population size N plus a residual amount of order $o\left(\frac{1}{N}\right)$. $V_i(t)$ satisfies a linear recurrence relation that uses differentiation of functions and the covariance of $\mu(t)$, as shown in Theorem 1 of [12] (see also [14]), and can be implemented efficiently using symbolic differentiation software packages. It is easy to see that the larger is N the smaller this additional factor gets. Essentially, the refined mean field takes not only the first moment (the mean) but also the second moment (variance) into consideration in the approximation. In [12] we have applied this discrete time refined mean field approximation on a number of examples ranging from the well-known epidemic model SEIR to wireless networks. Here we investigate its application to a novel model of the more complex gossip shuffle protocol.

A proof-of-concept implementation of both the classical and the refined mean field techniques and a discrete event simulator has been developed by one of the authors of the present paper in F# using the DiffSharp package [5] for symbolic differentiation. The results in this paper have been obtained using this implementation which can be found at [20].

4 Refined Mean Field Approximation of the Gossip Shuffle Protocol

The classical mean field model of the gossip protocol in [1], and aggregated versions thereof in [14], are based on the principle of decoupling of joint probabilities [19,7] and on a careful study of the pairwise probabilities of the various

possible outcomes of a shuffle between two gossip nodes. This model provides reasonable accuracy for systems with tens of thousands of nodes or more. However, discrete event simulation of this model for medium size systems shows that it does not respect important properties of the original gossip shuffle protocol, in particular the property that the new data element never gets lost from the system. We have found that this is caused by an inaccurate modelling of the *effects of coordination* between interacting nodes (see [14] for details).

We present a novel model in which (1) the system can never completely loose the inserted data element and (2) the model reflects the *effects* of the pairwise interaction between nodes satisfying basic properties of the original gossip shuffle protocol while still adhering to the principle of decoupling of joint probabilities. We distinguish the effects of a node getting a data element through *exchanging* it with another node–in which case the total number of replicas of the data element in the system remains the same–or through *replication*, i.e. the other node retains its copy of the data element and the global number of the data element in the system increases by one. With reference to Fig. 2, for what concerns point (1) above, we introduce the state PD to the model representing that there always is a gossip node in the network that possesses the data element.

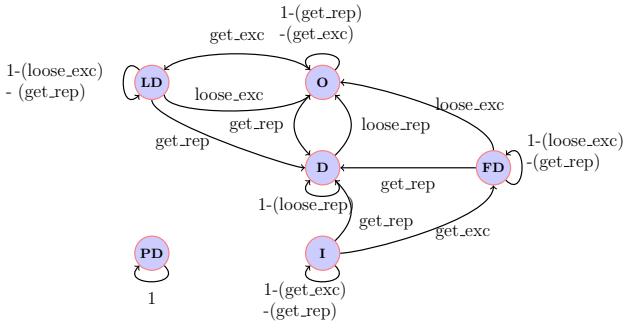

Fig. 2. Six-state model of an individual gossip node with rounds of length G_{max}.

To address point (2), we introduce states FD and LD to distinguish between the effect of interactions between gossip nodes. State FD represents the fact that the gossip node received the data element for the *first time* via an *exchange* of the data element with another node. State LD also represents the fact that the node received the data element via an *exchange*, but that it had already seen the data element in the past. Note that we can retrieve the total number of gossip nodes in the system that do *not* possess the data element as *the sum* of the nodes that are in states FD, LD, I and O because for each node in state FD (LD, resp.) there is a node in the network that just lost its data element in the synchronous shuffle with our current node. A gossip node can also get involved in an interaction in which the data element is replicated, i.e. a node

gives it to another one but also retains a copy itself, and one in which two nodes, both possessing the data element, interact and one of them looses its copy. Note that in this model it is not possible that both nodes loose their copy in a single interaction. The conditional probabilities of pairs of interacting nodes obtaining or loosing the data element can be expressed in terms of n (number of different data elements), c (size of the cache) and s (number of selected elements for exchange), as follows[5]:

$$\begin{aligned}
P(OD|DO) &= P(DO|OD) &&= \tfrac{s}{c} * \tfrac{n-c}{n-s} \\
P(OD|OD) &= P(DO|DO) &&= \tfrac{c-s}{c} \\
P(DD|OD) &= P(DD|DO) &&= \tfrac{s}{c} * \tfrac{c-s}{n-s} \\
P(OD|DD) &= P(DO|DD) &&= \tfrac{s}{c} * \tfrac{c-s}{c} * \tfrac{n-c}{n-s} \\
P(DD|DD) & &&= 1.0 - 2.0 * \tfrac{s}{c} * \tfrac{c-s}{c} * \tfrac{n-c}{n-s} \\
P(OO|OO) & &&= 1.0
\end{aligned}$$

The probability functions of the state transitions in the model below depend on m, i.e. the occupancy measure vector, the conditional probabilities, the 'no collision' probability noc, and G_{max}(see [14] for further details).

$$\begin{aligned}
\mathsf{get_exc}\,(m) &= 2 * \tfrac{G_{max}}{(G_{max}+1)^2}(m_D + m_{PD})P(OD|DO)\mathsf{noc} \\
\mathsf{get_rep}\,(m) &= 2 * \tfrac{G_{max}}{(G_{max}+1)^2}(m_D + m_{PD})P(DD|DO)\mathsf{noc} \\
\\
\mathsf{loose_exc}\,(m) &= 2 * \tfrac{G_{max}}{(G_{max}+1)^2}(m_O + m_I + m_{LD} + m_{FD})P(OD|DO)\mathsf{noc} \\
\mathsf{loose_rep}\,(m) &= 2 * \tfrac{G_{max}}{(G_{max}+1)^2}(m_D + m_{PD})P(DO|DD)\mathsf{noc}
\end{aligned}$$

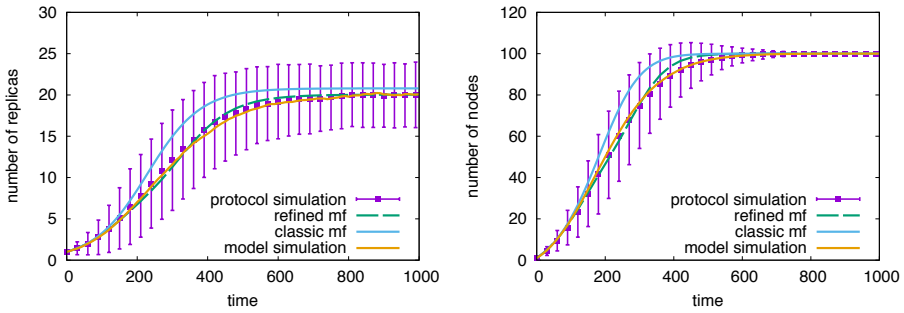

Fig. 3. Replication (left) and network coverage (right) of the data element in the network for $N = 100$ with initially 99 nodes in the I-state and 1 node in the PD-state for $G_{max} = 3$. Average of 500 simulation runs of both the model and Java simulations. Vertical bars show standard deviation for the Java simulation.

Fig. 3 shows the replication as sum of the number of nodes in states D and PD and the coverage as the sum of the number of nodes in D, PD, FD, LD and O for a network with $N = 100$, $n = 500$, $c = 100$ and $s = 50$ with initially one node

[5] $P(A'B'|AB)$ is the conditional probability of the state of an active-passive pair AB to have state $A'B'$ after their interaction, where $A, B, A', B' \in \{O, D\}$, see [1,2].

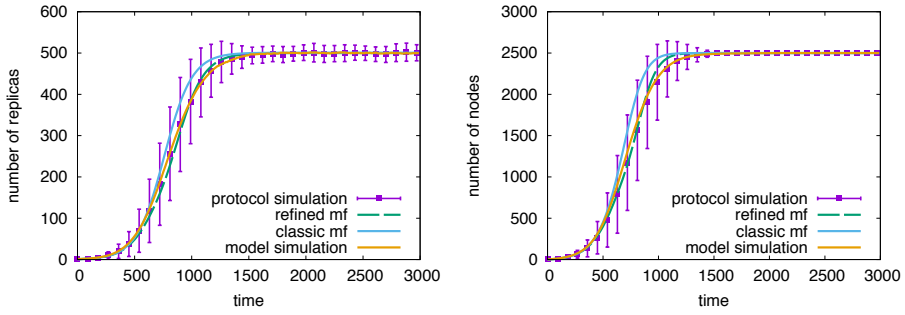

Fig. 4. Replication (left) and network coverage (right) of data element for $N = 2500$ with initially 2499 nodes in I and 1 in PD, for $G_{max} = 9$. Average of 500 simulation runs for both model and Java simulations. Vertical bars show standard deviation for the Java simulation.

in state PD and all the others in state I. Besides the classic and refined mean field approximations for the model in Fig. 2 and the Java simulation results of the actual shuffle protocol, Fig. 3 also shows the average of the model simulation. Note the good approximation of the simulation results by the refined mean field even in this very small network. Similarly good results have been found for a system with N=2,500 shown in Fig. 4. A first comparison of the (non-optimised) performance of the implementation in F# of the analysis for N=2,500, producing the results in Fig. 4, is: 0.5s (classic mean field); 25.5s (refined mean field[6]); 7m 1.4s (fast model simulation [19], 500 runs); 3h 42m 41.5s (Java simulation, 500 runs) on a MacBook Pro, Intel i7, 16GB.

5 Conclusions

We have developed a novel mean field model for the shuffle gossip protocol with which more accurate approximations for medium size gossip protocols can be obtained via refined mean field approximation techniques. This model respects key aspects of the protocol such as the effects of different kinds of interactions and the fact that a new data element cannot be lost by the system as a whole. Accurate approximation of medium size systems is important because many practical systems consist of many, but not a huge number of, components and simulation of such systems is still a resource consuming effort. A refined mean field approximation can provide very fast and accurate approximations.

[6] Recall that the mean field analyses times are *independent* of the size of the system.

References

1. Bakhshi, R.: Gossiping Models – Formal Analysis of Epidemic Protocols. Ph.D. thesis, Vrije Universiteit Amsterdam (January 2011), http://www.cs.vu.nl/en/Images/Gossiping_Models_van_Rena_Bakhshi_tcm210-256906.pdf
2. Bakhshi, R., Cloth, L., Fokkink, W., Haverkort, B.R.: Mean-field framework for performance evaluation of push-pull gossip protocols. Perform. Eval. 68(2), 157–179 (2011), https://doi.org/10.1016/j.peva.2010.08.025
3. Bakhshi, R., Gavidia, D., Fokkink, W., van Steen, M.: An analytical model of information dissemination for a gossip-based protocol. Computer Networks 53(13), 2288–2303 (2009), https://doi.org/10.1016/j.comnet.2009.03.017
4. Bakhshi, R., Gavidia, D., Fokkink, W., van Steen, M.: A modeling framework for gossip-based information spread. In: Eighth International Conference on Quantitative Evaluation of Systems, QEST 2011, Aachen, Germany, 5-8 September, 2011. pp. 245–254. IEEE Computer Society (2011), https://doi.org/10.1109/QEST.2011.39
5. Baydin, A.G., Pearlmutter, B.A., Radul, A.A., Siskind, J.M.: Automatic differentiation in machine learning: a survey. Journal of Machine Learning Research 18, 153:1–153:43 (2018), http://jmlr.org/papers/v18/17-468.html
6. Birman, K.: The promise, and limitations, of gossip protocols. Operating Systems Review 41(5), 8–13 (2007), https://doi.org/10.1145/1317379.1317382
7. Bortolussi, L., Hillston, J., Latella, D., Massink, M.: Continuous approximation of collective system behaviour: A tutorial. Perform. Eval. 70(5), 317–349 (2013)
8. Frei, R., Serugendo, G.D.M.: Advances in complexity engineering. International Journal of Bio-Inspired Computation 3(4), 199–212 (2011), https://doi.org/10.1504/IJBIC.2011.041144
9. Frei, R., Serugendo, G.D.M.: Concepts in complexity engineering. International Journal of Bio-Inspired Computation 3(2), 123–139 (2011), https://doi.org/10.1504/IJBIC.2011.039911
10. Gast, N., Gaujal, B.: A mean field approach for optimization in discrete time. Discrete Event Dynamic Systems 21(1), 63–101 (2011), https://doi.org/10.1007/s10626-010-0094-3
11. Gast, N., Houdt, B.V.: A refined mean field approximation. Proceedings of the ACM on Measurement and Analysis of Computing Systems 1(2), 33:1–33:28 (2017), https://doi.org/10.1145/3154491
12. Gast, N., Latella, D., Massink, M.: A refined mean field approximation of synchronous discrete-time population models. Perform. Eval. 126, 1–21 (2018), https://doi.org/10.1016/j.peva.2018.05.002
13. Gast, N., Latella, D., Massink, M.: A refined mean field approximation for synchronous population processes. In: Workshop on MAthematical performance Modeling and Analysis (MAMA 2018). pp. 30–32. ACM SIGMETRICS Performance Evaluation Review, ACM (2019)
14. Gast, N., Latella, D., Massink, M.: Refined mean field analysis of the gossip shuffle protocol – extended version – (2020), https://arxiv.org/abs/2004.07519, arXiv:2004.07519v1
15. Gavidia, D., Voulgaris, S., van Steen, M.: A gossip-based distributed news service for wireless mesh networks. In: Conf. on Wireless On demand Network Systems and Services (WONS). pp. 59–67. IEEE Computer Society (2006)
16. Jelasity, M.: Gossip. In: Serugendo, G.D.M., Gleizes, M.P., Karageorgos, A. (eds.) Self-organising Software - From Natural to Artificial Adaptation, pp.

139–162. Natural Computing Series, Springer (2011), https://doi.org/10.1007/978-3-642-17348-6_7
17. Latella, D., Loreti, M., Massink, M.: On-the-fly PCTL fast mean-field approximated model-checking for self-organising coordination. Sci. Comput. Program. 110, 23–50 (2015), https://doi.org/10.1016/j.scico.2015.06.009
18. Latella, D., Loreti, M., Massink, M.: Flyfast: A mean field model checker. In: Legay, A., Margaria, T. (eds.) Tools and Algorithms for the Construction and Analysis of Systems - 23rd International Conference, TACAS 2017, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2017, Uppsala, Sweden, April 22-29, 2017, Proceedings, Part II. Lecture Notes in Computer Science, vol. 10206, pp. 303–309 (2017), https://doi.org/10.1007/978-3-662-54580-5_18
19. Le Boudec, J., McDonald, D.D., Mundinger, J.: A generic mean field convergence result for systems of interacting objects. In: Fourth International Conference on the Quantitative Evaluaiton of Systems (QEST 2007), 17-19 September 2007, Edinburgh, Scotland, UK. pp. 3–18. IEEE Computer Society (2007)
20. Massink, M.: Refined mean field F# implementation and gossip shuffle model, https://github.com/mimass/RefinedMF
21. Pianini, D., Beal, J., Viroli, M.: Improving gossip dynamics through overlapping replicates. In: Lluch-Lafuente, A., Proença, J. (eds.) Coordination Models and Languages - 18th IFIP WG 6.1 International Conference, COORDINATION 2016, Held as Part of the 11th International Federated Conference on Distributed Computing Techniques, DisCoTec 2016, Heraklion, Crete, Greece, June 6-9, 2016, Proceedings. Lecture Notes in Computer Science, vol. 9686, pp. 192–207. Springer (2016), https://doi.org/10.1007/978-3-319-39519-7_12
22. Voulgaris, S., Jelasity, M., van Steen, M.: A robust and scalable peer-to-peer gossiping protocol. In: Moro, G., Sartori, C., Singh, M.P. (eds.) Agents and Peer-to-Peer Computing, Second International Workshop, AP2PC 2003, Melbourne, Australia, July 14, 2003, Revised and Invited Papers. Lecture Notes in Computer Science, vol. 2872, pp. 47–58. Springer (2003), https://doi.org/10.1007/978-3-540-25840-7_6