Democratizing Data Analytics

Crowd-sourcing Decentralized Collective Measurements

Evangelos Pournaras, Edward Gaere, Renato Kunz, Atif Nabi Ghulam
Professorship of Computational Social Science, ETH Zurich, Zurich, Switzerland

dias-net.org

1 Challenge

How data consumers can accurately estimate aggregation functions having as input shared data of data suppliers?

PRIVACY-PRESERVING DECENTRALIZED NETWORK

DATA *SUPPLIERS*: THEY *SHARE* DATA

DATA *CONSUMERS*: THEY *AGGREGATE* DATA

2 Applicability

Participatory Smart City Crowd-sensing

TOTAL ENERGY CONSUMPTION

AVERAGE TRAFFIC FLOW

AVERAGE NOISE POLUTION

VOTING & SELF-GOVERNACE

REPUTATION & RATING MEASUREMENTS

3 Accuracy vs. Decentralization

Adaptive Aggregates

System Dynamics

1 *Changing data*

2 *Join & leaves*

3 *Failures*

5 DIAS: Dynamic Intelligent Aggregation Service

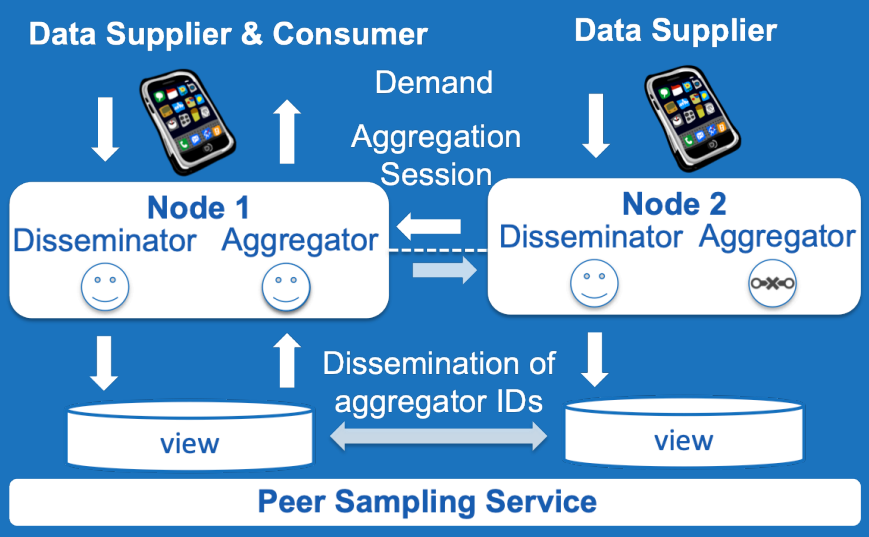


Fig. 1. Joining/leaving process: A data consumer connects at Node 1. It is detected by Node 2 via gossip communication. Three extreme join/leave profiles are shown in (a) and how DIAS adapts the estimation of the total Smart Grid consumption in a network of 3000 nodes as shown in (b).

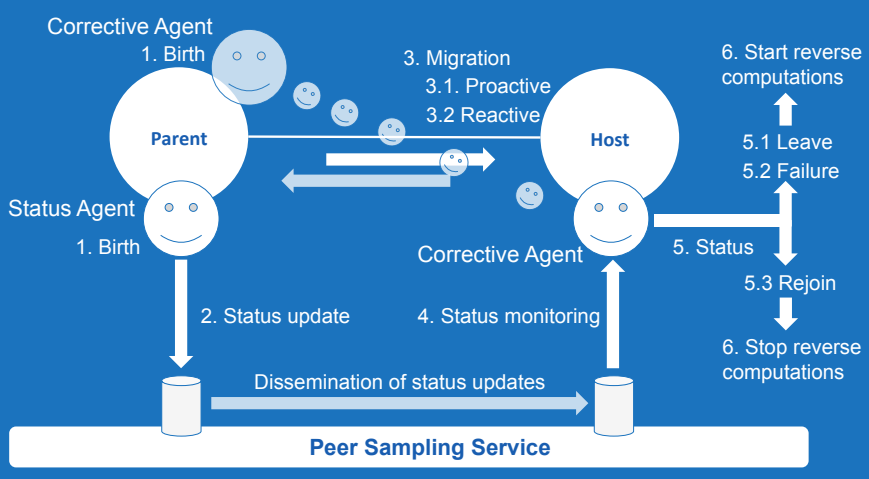
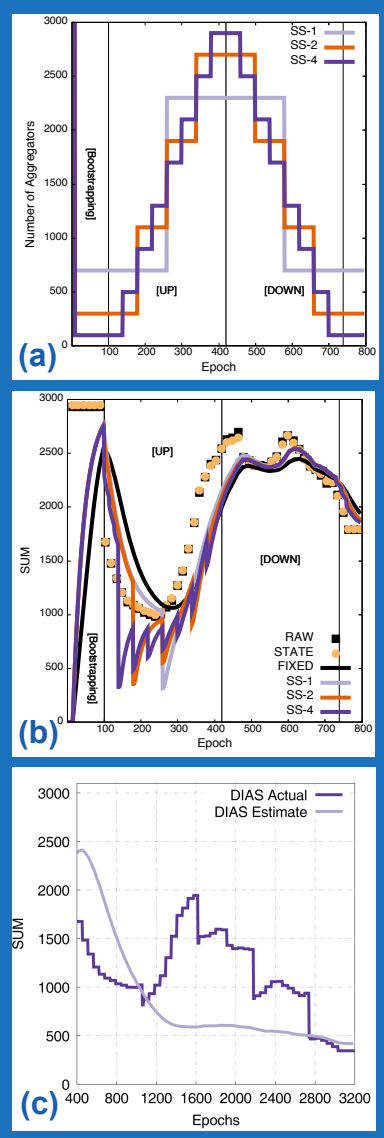


Fig. 2. Fault-tolerance process: A corrective agent migrates proactively (to mitigate failures) or reactively (to mitigate leaving) from its parent to another host. A status agent at the parent sends updates to signal its connectivity. Based on these updates, the corrective agent can initiate reverse computations that improve the accuracy of the aggregates. An extreme example with 80% of the data suppliers failing is shown in (c) in which the threashold of the corrective agent before starting the reverse computations is set to 250 epochs.



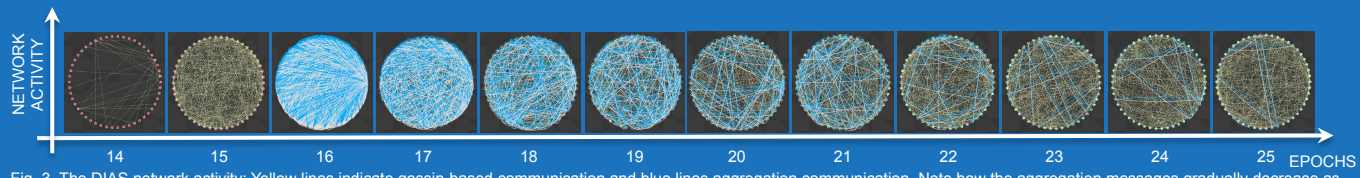


Fig. 3. The DIAS network activity: Yellow lines indicate gossip-based communication and blue lines aggregation communication. Note how the aggregation messages gradually decrease as the aggregate estimates converge to the actual values indicated by the red nodes turning into green ones.

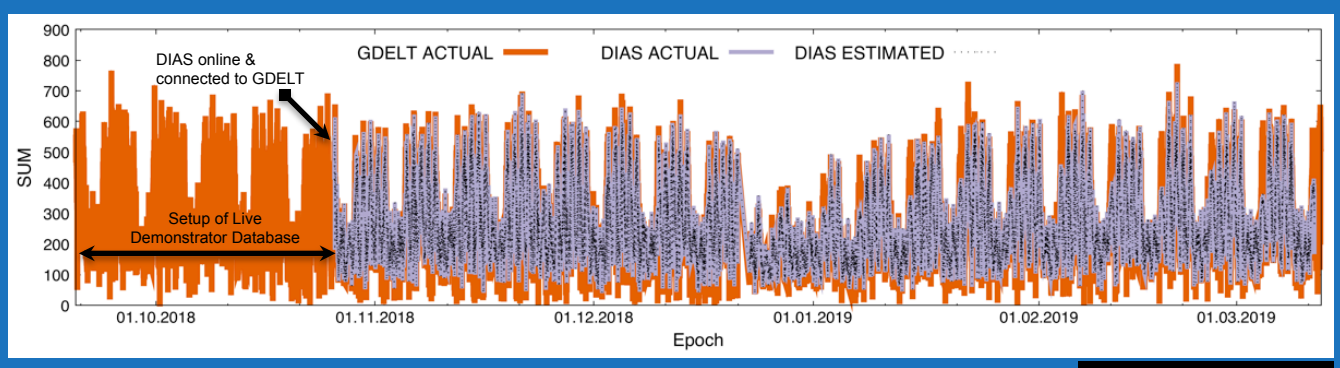


Fig. 4. DIAS-GDELT TRL-6 live demonstrator in operation for 4 months exchanging million of messages to adapt the aggregates to the actual values. These values represent the total number of GDELT news events from 28 countries. The news events can be accessed via the GDELT API and they update every 15 minutes.

GDELT API
gdeltproject.org

4 Design Solution

Information dissemination: gossiping

Distributed memory: Bloom filters

Fault-tolerance: agent migration

6 Conclusion

TRL-6 feasibility & proof of concept

Accuracy under extreme dynamics

Data analytics by citizens, for citizens

7 References

- E. Pournaras, J. Nikolic, A. Omerzel, and D. Helbing, "*Engineering democratization in internet of things data analytics*," in 2017 IEEE 31st International Conference on Advanced Information Networking and Applications (AINA). IEEE, 2017, pp. 994–1003.
- E. Pournaras and J. Nikolic´, "*Self-corrective dynamic networks via decentralized reverse computations*," in 2017 IEEE International Conference on Autonomic Computing (ICAC). IEEE, 2017, pp. 11–20.
- E. Pournaras and J. Nikolic, "*On-demand Self-adaptive Data Analytics in Large-scale Decentralized Networks*," in the Proceedings of the 16th IEEE International Symposium on Network Computing and Applications (NCA). IEEE, 2017, pp. 1-10
- http://dias-net.org/dias-gdelt-live
- https://github.com/epournaras/DIAS
- https://github.com/epournaras/DIAS-GDELT
- https://github.com/epournaras/DIAS-Documentation

Contributors: Jovan Nikolic - Nursultan Jubatyrov - Srivatsan Yadhunathan - Pablo Velasquez - Aleš Omerzel - Varin Buff - Melika Ayoughi