MSc Thesis

What topics should be discussed / described in the thesis (these are merely some suggestions!).

1. introduction – preliminaries, description of the studied problem and literature survey
   1.1. preliminaries – notations used, some mathematical / theoretical definitions, main theorems etc.
   1.2. a short paragraph about wireless sensor networks (WSNs) / algorithmics of WSNs.
   1.3. distributed data aggregation (incl. large datasets / data streams) / data aggregation in WSNs
      - formulation of the problem
      - the idea of probabilistic counting and application of probabilistic counters to aggregation in large datasets
      - description of the algorithm HISTMEAN from [CG18] and the approach to averaging time-series data from [NWSC20]
      - brief mention of some other important algorithms for distributed data aggregation known from literature
   1.4. simulations of distributed computing systems – what are the key issues, some examples of existing solutions / distributed computing platform etc.
2. description of the implemented simulator (rather concise)
   2.1. motivations (*"Why I've designed and implemented a custom solution instead of using one of the existing simulators?"*) and project objectives
   2.2. technical aspects of the implementation (short description of used technologies and design patterns), architecture, design objectives etc.
   2.3. implemented features, main use cases, what it offers and how it works, possible directions for further extension / improvements
   2.4. comparison with existing solutions (if there will be enough time)
3. averaging time-series data – applications of the implemented simulator to the study of selected problems in distributed aggregation of large datasets in WSNs by the example of building distributed data sketches and averaging of time-series data in WSNs
   3.1. motivations and description of considered scenarios.
   3.2. description of considered variants of analyzed algorithms (some modifications of the approach from [NWSC20]).
   3.3. quantities of interest – which parameters / properties of considered algorithms will be analyzed and why?
      - estimator's precision
      - communication complexity (message complexity) for selected families of graphs – e.g. line graphs L_n, grid graphs $G_{k\times k}$, random geometric graphs $\mathcal{R}_{n,r}$, d-regular graphs, "grid-of-cliques", etc. $\rightarrow$ choose up to 2–3 examples
4. simulation results – experimental evaluation of estimator's precision and message complexity
   4.1. presentation (tables, graphs, charts, ...) and discussion on the obtained results of simulations for different scenarios

- different variants of the algorithm
- different input data (various distributions of the data, different network sizes n)
- parameter 1: precision of the obtained estimates of the average
- parameter 2: message complexity for certain network topologies

4.2. analysis of the results of performed experimental research, drawing some conclusions and stating hypotheses

5. theoretical analysis of some selected properties of the considered protocols for averaging and building distributed data sketches
   - based on the obtained simulation results we should choose some quantity, e.g. estimator's precision, state some hypotheses and try to prove some theoretical results in the assumed formal model
   - it suffices to show one – two sample results
   - e.g. something like "In the basic variant A of the averaging algorithms, the estimator's probabilistic error grows as $t \to \infty$". and "In the strategy implemented in the variant B, the estimator's probabilistic error depends on the parameter R and for constant R tends to 0 in expectation as the numbers of counters L grows."
   - what will be analyzed – we'll take a look at the experimental results and decide, what seems to be achievable in 2-3 weeks (we don't have a lot of time)
6. Summary
   - summary of the obtained experimental and theoretical results
   - conclusions
   - open problems and future work

Considered scenarios

The network will be modeled by a simple, connected and undirected graph $G = (V, E)$ with $|V| = n$ vertices representing the stations and $|E| = m$ edges corresponding to bidirectional links between nodes. Two stations $u, v \in V$ can communicate directly if and only if $\{u, v\} \in E$...

Assumptions on the considered theoretical model:

- **We'll discuss the details during the next call**
- each station observes a stream of data $(s_t^{(v)})_{t \geq 0}$ arriving in consecutive time instants $t = 0, 1, \ldots$ – like in [NWSC20]
- our goal is to calculate the estimations $(avg_t)_{t \geq 0}$ in a fully distributed manner, where

$$avg_t \stackrel{\text{df}}{=} \frac{\sum_{v \in V} s_t^{(v)}}{n}$$

- we will consider certain modifications of the method proposed in [NWSC20], which extends the approach based on approximate histograms introduced in [CG18].
- communication model – like in [CG18] (different model than in [NWSC20]), i.e. broadcast-based communication instead of gossip-based (**this is only my suggestion – you can choose what you prefer)**

- measures of precision – I suggest implementing and calculating both measures of precision, i.e. $\mathrm{err}(\mathrm{wm}(\vec{C}_L, \mathrm{wm}\vec{H})$ introduced in [CG18] and η from [NWSC20]. We'll see what we'll get. In the thesis we can e.g. briefly mention, that the precision can be defined in different ways, depending on the specific applications, considered scenarios etc.

- considered strategies (some modification of the approach of counting "leavers" and "joiners" for each interval):

  - variant 0 – for each t we independently run the HISTMEAN algorithm with fixed L and K
  - variant 1 – the strategy from [NWSC20] – probabilistic counters for "leavers" $C_{L,i}^{leave}$ and "joiners" $C_{L,i}^{join}$ for each interval I_i, $i \in \{1, \ldots, K\}$ (the number of leavers and joiners will grow as $t \to \infty \Rightarrow$ I suppose that the error related to probabilistic counters will also grow and the precision will be lost for large t, but this is only a conjecture and this may not be the case – we should investigate this)
    Q: In [NWSC20] there was a flag r_{flag}, which affects the exchange of messages – should we embed this mechanism in the broadcast-based communication or not? We'll discuss it during the next call.
  - variant 2 – "reset" the counters every R^{th} round (e.g. $joiners \leftarrow \max\{joiners - leavers, 0\}$, $leavers \leftarrow 0$)
  - variant 3 – every R^{th} round "reset" the counters by running the HISTMEAN algorithm from scratch

- in each case we will calculate the estimator's precision (as a function of the time t) for different network sizes n (e.g. $n = 10,\ 100,\ 1000,\ 5000$)

- distribution of the input data

  - in the basic scenarios we assume that the data are independent (at least there is no spatial dependence)
  - uniform distribution over some interval $[m, M]$ (e.g. over the unit interval)
  - data for which the stations "will change intervals" rarely
  - data for which the changes of intervals will be very frequent

- for some scenarios and network topologies calculate the message complexity (the number of exchanged messages)

Future work

- data are generated with different intensities, some stations update their values more frequently than the others

- data streams are not time-homogeneous – the distribution of the data $(s_t^{(v)})_{t \geq 0}$ for a given station v may vary in time

- spatial and temporal dependencies of data

- communication errors – broken links, nodes failures

References

[CG18] Jacek Cichoń and Karol Gotfryd. Average counting via approximate histograms. *ACM Trans. Sen. Netw.*, 14(2):8:1–8:32, March 2018.

[NWSC20] Saptadi Nugroho, Alexander Weinmann, Christian Schindelhauer, and Andreas Christ. Averaging emulated time-series data using approximate histograms in peer to peer networks. In Fernando De La Prieta et al., editor, *Highlights in Practical Applications of Agents, Multi-Agent Systems, and Trust-worthiness. The PAAMS Collection*, pages 339–346, Cham, 2020. Springer International Publishing.