Systematic Literature Review Protocol: Data Dissemination in Large-cardinality Social Graphs

Tymur Maryokhin
Linnaeus University,
Växjö, Sweden

June 15, 2015

1 Change record

<table>
<thead>
<tr>
<th>Document History</th>
</tr>
</thead>
<tbody>
<tr>
<td>Version 0.1</td>
</tr>
<tr>
<td>Version 0.2</td>
</tr>
<tr>
<td>Version 0.3</td>
</tr>
<tr>
<td>Version 1.0</td>
</tr>
</tbody>
</table>

2 Background

In the problem area of real-time message delivery there are very strict requirements on what is considered to be "real time" [1]. These requirements are especially hard to uphold for large-scale systems which social media applications usually are. As the overall amount of Internet traffic continues to increase [2], the load on social media applications also increases, meaning that new problems arise over time.

One of these problems is the clusterization of social graphs, the concept that interactions in social media applications are built on. Clusterization happens when select producers start to attract many consumers, forming clusters, or large-cardinality local social graphs. Therefore, even though the requirements for data dissemination through all the social graphs are the same, the laws governing message delivery through small and large cardinality graphs are completely different. While these laws have not yet been clearly defined, their effect on the architecture of social media applications is great. As a result, the problem was recognized by companies in industry [3], and multiple solutions were proposed and developed [4]–[6].

This study applies the Systematic Literature Review (SLR) methodology [7] to investigate the challenges of data dissemination in large-cardinality social graphs. Since no past SLR has been found regarding the topic, it was chosen to conduct a systematic review as a reliable tool [8] aimed at identifying gaps in previous research, summarizing existent evidence and positioning future research.

3 Research questions

To ensure that the study will be broad, covering all the topics related to the field, the following research questions have been formed:

[SLRQ1] What problems do large-cardinality social graphs pose for existing large-scale social network applications?

[SLRQ2] What is the current practice of dealing with large-cardinality social graphs in the enterprise?

Along with answering the research questions, the end goal of this SLR is to:
• Define research areas which should be studied to analyze the problem of data dissemination in large-cardinality graphs.

• Conduct a systematic review of the literature found in the related research areas regarding large-cardinality graphs.

• Select the most relevant studies to be reviewed in greater detail.

• Collect empirical data offered to the public by companies in industry dealing with the same problem.

• Provide an up-to-date picture of the state of research in this field.

• Offer a direction for further research in this field.

4 Search strategy

4.1 Resources

Journals


Electronic databases

Google Scholar, ISI Web of Science, ACM Digital Library, CiteSeerX, EBSCOhost, IEEExplore.

Conference proceedings


Online resources

High Scalability, Yahoo! Labs, online articles and multiple engineering blogs (Facebook, Twitter, Instagram, Tumblr, Flickr, etc.) will be referenced for up-to-date empirical data collected by companies in industry.

4.2 Search process

The search process will primarily consist of performing full text searches on the sources outlined above. The following keywords have been identified as the most relevant: data dissemination, message delivery, large-cardinality social graph, large scale, feed following, feed construction, materialized views, social network analysis, community structure detection, graph theory, database theory. The searches will include various combinations of these keywords. After a primary search has been concluded, multiple secondary searches will be performed to ensure that all relevant material was captured.

4.3 Validation

In the process of a previously conducted systematic mapping study, several publications were identified as being the most relevant to the research area at hand:


These papers were identified during pilot searches against electronic databases, such as Google Scholar, IEEExplore and ACM digital library. The results of these pilot searches were then used to validate the inclusiveness of the selected search terms. The quantity and the quality of the returned studies were then evaluated to develop a schedule (section 9), estimating how long the whole SLR process should take.

5 Selection criteria

5.1 Inclusion criteria

- Publications whose main focus is on analyzing data dissemination in social graph based networks.
- Publications which discuss approaches already applied in industry will be prioritized over publications discussing potential approaches, but neither will be excluded.
- When there are several studies reporting the same topic, only the most recent will be taken.
- When there is empirical data provided by the same source in different periods of time, the most recent will be taken.
- Date of publication or non-academia based origin of reports are not detrimental to inclusion.
- Gray literature (technical reports and programming conference proceedings) will be accepted if relevant.

5.2 Exclusion criteria

- Studies whose main focus is not related to the research areas targeted in this study. The areas of research in the beginning are: social network analysis, graph theory, database theory. The research scope might be widened or narrowed in the process and will be reflected in the protocol accordingly.
- Papers where only the abstract or summary, but not the full text, is available.
- Publications that provide empirical data collected from a small-scale network (less than 100'000 nodes).
- Outdated studies based on approaches which have been proven to not work at large scale in industry.
- Papers that discuss a purely theoretical approach, but do not have empirical data to prove the point.
- Publications not written or translated to English.

5.3 Selection process

The selection process will be performed by one researcher (the author) in an iterative manner. After fixed intervals the selection process will be reapplied to the selected studies to verify the correctness of the selection. To additionally verify the selection process, a second researcher (master thesis supervisor) will perform random sampling of the selected studies to identify the validity of selection. The selection process is defined in two phases:
1. Publications go through preliminary selection based on their title and abstract. Studies that are clearly irrelevant are not processed further.

2. Publications selected during the previous phase are analyzed in depth in search for relevant content. Publications that pass this selection phase should have information which is valuable for future extraction and analysis.

5.4 Quality assessment

Each selected publication will need to be assessed for its quality. The goal of the quality assessment is to evaluate whether the study at hand provides a valuable source of information to the research or not. Therefore, criteria have been developed which will be used to evaluate the contribution of each publication to the SLR. The criteria are as follows:

1. Is the paper based on previous research or on empirical data gathered in a production environment?
2. Is the architecture approach/design discussed in the work applicable to a large-scale production environment?
3. Is the aim of research clearly stated?
4. Was the data collected in a way that addressed the research question and did not distort the findings?
5. Was the data analysis sufficient?
6. Is the context, in which the research was carried out, thoroughly described?
7. Are the results of the research clearly stated?
8. Is there data or other research papers which prove the claims made in this work?
9. Is the research based on up-to-date data?
10. Is the study of value for research and practice?

The first two criteria will be applied to exclude papers which are not scientific or are not applicable to the domain of large-scale applications. The body of each work will then be analyzed to evaluate if the aim of research is clearly stated, the data was collected and analyzed correctly, and that the results were properly documented. In the domain of large scale, things tend to change very quickly, and therefore each study needs to be assessed for not being based on outdated data. After that, the results of each paper will be compared to the results of previously found papers for additional validity.

6 Data extraction

6.1 Data collection

The data extraction process will be performed by one researcher (the author). As with the study selection, a second assessor (master thesis supervisor) will perform random sampling of the extracted data to evaluate the relevance of the collected information. The following information is to be extracted from the selected studies:

• Reason for acceptance of the study for the SLR.
• Study type (i.e. conference paper, online article, journal publication, etc.).
• Study goals and objectives.
• Methodology of the study (observational, controlled experiment, case study, etc.)
6.2 Data synthesis

Until all the data has been extracted, the data synthesis strategy is subject to change. It is, however, anticipated that descriptive synthesis will be chosen. Data will be organized in a tabular fashion consistent with the review questions. Data dissemination will take form in a narrative description of the results presented as the final report.

7 Study limitations

One of the limiting factors of this research is that the author of the study does not have control over the resources or electronic databases which will be used as sources for the SLR. As there is no influence on the results that are returned by these source, there is no guarantee that over time it would be possible to replicate the exact research described in this protocol.

Another influential factor, which can be considered a limitation, is personal bias. As it is not possible to avoid the impact that each person’s background and experience causes, the selection criteria of the studies might have been different depending on the researcher at hand. The risk of subjectivity will, however, be minimized by random sampling and reviews done on a continuous basis by a person other than the author, namely by the supervisor of the master thesis project.

Last but not least, it is important to note the subjectivity of the collected empirical data. Due to the specifics of the research area of large scale applications, the author does not possess the resources to perform a series of controlled experiments on a self-hosted application. Therefore, data collected by companies in industry and provided to the public will be collected and used in the research as is. To increase the validity of the found empirical data, cross-reference checks will be used whenever possible.

8 Validation of the protocol

Validation of the protocol has been performed on version 0.2 of the protocol. Revision was performed by Ola Petersson (Associate Professor at Linnaeus University). After corrections noted by the expert review were taken into account, version 0.3 was drafted.

9 Schedule

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
<th>Deliverables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 1</td>
<td>Develop SLR protocol</td>
<td>n/a</td>
</tr>
<tr>
<td>Week 2</td>
<td>Revise SLR protocol</td>
<td>Finalized protocol</td>
</tr>
<tr>
<td>Weeks 3-6</td>
<td>Search for publications</td>
<td>List of publications</td>
</tr>
<tr>
<td>Week 7</td>
<td>Publication selection</td>
<td>List of selected publications</td>
</tr>
<tr>
<td>Week 8</td>
<td>Data extraction</td>
<td>Report notes</td>
</tr>
<tr>
<td>Weeks 9-10</td>
<td>Data analysis</td>
<td>n/a</td>
</tr>
<tr>
<td>Weeks 11-14</td>
<td>Report write-up</td>
<td>Finalized report</td>
</tr>
</tbody>
</table>
References


