



**University
of Victoria**

Graduate Studies

Notice of the Final Oral Examination
for the Degree of Doctor of Philosophy

of

MICHEAL SIMPSON

MSc (University of Victoria, 2014)

BSc (University of Victoria, 2011)

**“Scalable Algorithms for Misinformation
Prevention in Social Networks”**

Department of Computer Science

Friday, December 7, 2018

10:00 A.M.

Engineering and Computer Science Building
Room 468

Supervisory Committee:

Dr. Alex Thomo, Department of Computer Science, University of Victoria (Co-Supervisor)

Dr. Venkatesh Srinivasan, Department of Computer Science, UVic (Co-Supervisor)

Dr. Michael Adams, Department of Electrical and Computer Engineering, UVic (Outside Member)

External Examiner:

Dr. Yong Gao, Department of Computer Science, UBC Okanagan

Chair of Oral Examination:

Dr. Reuven Gordon, Department of Electrical and Computer Engineering, UVic

Dr. David Capson, Dean, Faculty of Graduate Studies

Abstract

This thesis investigates several problems in social network analysis on misinformation prevention with an emphasis on finding solutions that can scale to massive online networks. In particular, it considers two problem formulations related to the spread of misinformation in a network that cover the elimination of existing misinformation and the prevention of future dissemination of misinformation. Additionally, a comprehensive comparison of several algorithms for the feedback arc set (FAS) problem is presented in order to identify an approach that is both scalable and computes a lightweight solution. The feedback arc set problem is of particular interest since several notable problems in social network analysis, including the elimination of existing misinformation, crucially rely on computing a small FAS as a preliminary.

The elimination of existing misinformation is modelled as a graph searching game. The problem can be summarized as constructing a search strategy that will leave the graph clear of any misinformation at the end of the searching process in as few steps as possible. Despite the problem being NP-hard, even on directed acyclic graphs, this thesis presents an efficient approximation algorithm and provides new experimental results that compares the performance of the approximation algorithm to the lower bound on several large online networks. In particular, new scalability goals are achieved through careful algorithmic engineering and a highly optimized preprocessing step.

The minimum feedback arc set problem is an NP-hard problem on graphs that seeks a minimum set of arcs which, when removed from the graph, leave it acyclic. A comprehensive comparison of several approximation algorithms for computing a minimum feedback arc set is presented with the goal of comparing the quality of the solutions and the running times. Additionally, careful algorithmic engineering is applied for multiple algorithms in order to improve their scalability. In particular, two approaches that are optimized (one greedy and one randomized) result in simultaneously strong performance for both feedback arc set size and running time.

The experiments compare the performance of a wide range of algorithms on a broad selection of large online networks and reveal that the optimized greedy and randomized implementations outperform the other approaches by simultaneously computing a feedback arc set of competitive size and scaling to web-scale graphs with billions of vertices and tens of billions of arcs. Finally, the algorithms considered are extended to the probabilistic case in which arcs are realized with some fixed probability and a detailed experimental comparison is provided.

Finally, the problem of preventing the spread of misinformation propagating through a social network is considered. In this problem, a “bad” campaign starts propagating from a set of seed nodes in the network and the notion of a limiting (or “good”) campaign is used to counteract the effect of misinformation. The goal is to identify a set of k users that need to be convinced to adopt the limiting campaign so as to minimize the number of people that adopt the “bad” campaign at the end of both propagation processes. *RPS* (Reverse Prevention Sampling), an algorithm that provides a scalable solution to the misinformation prevention problem, is presented. The theoretical analysis shows that *RPS* runs in $O((k + l)(n + m) \left(\frac{1}{1-\gamma}\right) \log n / \epsilon^2)$ expected time and returns a $(1 - 1/e - \epsilon)$ approximate solution with at least $1 - n^{-1}$ probability (where γ is a typically small network parameter). The time complexity of *RPS* substantially improves upon the previously best-known algorithms that run in time $\Omega(mnk \cdot \text{POLY}(\epsilon^{-1}))$. Additionally, an experimental evaluation of *RPS* on large datasets is presented where it is shown that *RPS* outperforms the state-of-the-art solution by several orders of magnitude in terms of running time. This demonstrates that misinformation prevention can be made practical while still offering strong theoretical guarantees.