ERCIM "ALAIN BENSOUSSAN"
FELLOWSHIP PROGRAMME

ERCIM
European Research Consortium
for Informatics and Mathematics

Scientific Report

First name / Family name	Nima Dokoohaki
Nationality	Iranian
Name of the *Host Organisation*	Swedish Institute of Computer Science (SICS Swedish ICT)
First Name / family name of the *Scientific Coordinator*	Daniel Gillblad
Period of the fellowship	15/07/2014 to 15/07/2015

I – SCIENTIFIC ACTIVITY DURING YOUR FELLOWSHIP

1) **Privacy-preserving mobility data mining and analysis**
A major part of my work was contributing to developing a privacy-preserving framework for mobility data analysis and publication. With this direction robust and privacy-preserving mobility modelling techniques using cellular network data was proposed. The specific aim was extracting common patterns of mobility throughout a metropolitan area from large-scale cellular datasets in a transportable and presentable format that can be published in an untraceable and privacy-preserving manner. To this end differential privacy was proposed and used, as it provides formal guarantees for individual data not to be traceable from aggregates, as well as privacy metric for controlling how much perturbation can be added to data in order to assure publisher of desired level of privacy and in-distinguishability. The approach was applied on 2014 dataset of Data for Development (D4D) challenge, and the initial results were published in [1]. Extended results were later on published in [2].
2) **Leveraging social media data for predictive behaviour modelling**
As continuation of my doctoral research, I extended my hands on experience with crawling and mining large-scale social media data, as explained in [4]. Focusing specifically on Twitter, we analysed a 2014 Twitter dataset crawled from Sweden mainly for the aim of electoral voting behaviour analysis. We analysed both content and social features of tweets of both Swedish politicians and their respective party accounts. We used both supervised and

unsupervised topic models for extracting agenda related topics and hashtags, this part will be published in [7]. For understanding the inherent social features underlying their interactions we used unsupervised link prediction, specifically SALSA algorithm was chosen and used along with personalized PageRank for popularity estimation. We identified high correlation between estimated popularities and official voting outcome. This part of results is explained in [3].

3) **Understanding and leveraging user profiles for Cache optimization**
Understanding, discerning and leveraging user behaviour patterns in mobile communication networks could have visible impact on how network providers deliver their content across mobile networks to their respective customers in an efficient manner. By unravelling such behaviour patterns, networks could optimize their caching mechanisms to adapt to global access patterns on a macroscopic level, or individual access patterns on a microscopic level. As a result, during the course of my research I have been working on how to leverage large-scale data analysis techniques for sampling and aggregating the access patterns in a country wide 3G network dataset, to understand temporal access patterns of users to multimedia data such as video and audio content, which will be explained in [5]. This was achieved by fitting the aggregates on a time series with respect to various access types of multimedia content and also variations of temporal accesses. In a later project we took a more personalized aim at understanding access patterns of each user to multimedia content. Focusing on a video on demand services (VoD) dataset, we proposed for leveraging supervised link prediction in order to find similar temporal patterns of access to online video and understanding implications of using the predictive link attributes on the graph for personalized caching. We have experimented with both supervised and unsupervised predictors. Given the long tail effect inherent in data Preferential Attachment has been the main predictor entailing the access behaviour. Existing results will be explained in [6].

4) **Scientific community contributions:**
I have been scientific reviewer for following journals:
Elsevier Information Systems,
Springer Knowledge and Information Systems,
IEEE transactions on human-machine systems,
KSII transactions on internet and information systems
European conference on Information Systems (ECIS 2015)
I have served as committee member of following venues:
IEEE international conference on social computing and networking (SocialCom'14),
2015 international workshop on mobile social networking and computing (MSNCom-2015),
2015 International workshop on Intelligent Exploration of Semantic Data (IESD 2015)

II – PUBLICATION(S) DURING YOUR FELLOWSHIP

Accepted publications:

1) Travel Demand Analysis with Differentially Private Releases. David Gundlegard, Clas Rydergren, Jaume Barcelo, Nima Dokoohaki, Olof Görnerup, Andrea Hess.
In proceedings of the Data for Development (D4D) Senegal challenge track, presented at NetMob

2015. (Published in book of abstracts NetMob conference website)
Abstract: When analyzing large quantities of human mobility traces, the aspects of sensitivity of traces to be analyzed, and the scale at which such analysis can be accounted for is of high importance. The sensitivity implies that identifiable information must not be inferred from the data or any analysis of it. Thus, prompting the importance of maintaining privacy during or post-analysis stages. We aggregate the raw data with the goal to retain relevant information while at the same time discard sensitive user specifics, through site sequence clustering and frequent sequence extraction. These techniques have at least three benefits: data reduction, information mining, and anonymization. Further, the paper reviews the aggregation techniques with regard to privacy in a post-processing step. The approaches presented in the paper for estimation of travel demand and route choices, and the additional privacy analysis, build a comprehensive framework usable in the processing of mobile phone data for transportation planning.

2) Privacy-preserving mining of frequent routes in cellular network data
Olof Görnerup, Nima Dokoohaki, Andrea Hess
In proceedings of the 14th IEEE international Conference on Trust, Security and Privacy in Computing and Communications (IEEE TrustCom-15), Helsinki, Finland 20-22 August, 2015.
Abstract: Cellular networks generate a wealth of mobility data that may be applied in numerous application areas such as traffic and transport management, urban planning and crisis management. Due to the sheer size of network data, and since it can contain sensitive information this potential also comes with great technical challenges with regard to information extraction, data reduction and privacy. In this paper we simultaneously address these three challenges with respect to the specific task of mining frequent routes of terminals and, in extension, people and vehicles. The proposed approach is a pipeline where raw cell data is segmented into sequences that, in turn, are aggregated into groups of similar sequences using a scalable distributed clustering algorithm based on locality-sensitive hashing. Acquired aggregate statistics are then perturbed using an existing differential-privacy framework in order to ensure that sensitive information is not released. This compound approach has been evaluated with respect to cluster quality, data reduction and privacy protection on coarse empirical call detail record data, as well as on more fine-grained synthetic handover data simulated from measured GPS traces.

3) Predicting Swedish Elections with Twitter: A Case for Stochastic Link Structure Analysis
Nima Dokoohaki, Filippia Zikou, Daniel Gillblad, Mihhail Matskin.
In 4th International Workshop on Social Network Analysis in Applications (SNAA). To appear in proceedings of 2015 IEEE/ACM International Conference on Advances in Social Network Analysis and Mining (ASONAM 2015).
Abstract: The question that whether Twitter data can be leveraged to forecast outcome of the elections has always been of great anticipation in the research community. Existing research focuses on leveraging content analysis for positivity or negativity analysis of the sentiments of opinions expressed. This is while analysis of link structure features of social networks underlying the conversation involving politicians has been less looked. The intuition behind such study comes from the fact that density of conversations about parties along with their respective members, whether explicit or implicit, should reflect on their popularity. On the other hand, dynamism of interactions can capture the inherent shift in popularity of accounts of politicians. Within this manuscript we present evidence of how a well-known link prediction algorithm, can reveal an authoritative structural link formation within which the popularity of the political accounts along with their neighbourhoods, shows strong correlation with the standing of electoral outcomes. As evidence, the public time-lines of two electoral events from 2014 elections of Sweden on Twitter have been studied. By distinguishing between member and official party accounts, we report that even using a focus crawled public dataset, structural link popularities bear strong statistical similarities with vote outcomes. In addition we report strong ranked dependence between standings of selected politicians and general election outcome, as well as for

official party accounts and European election outcome.

4) Diversifying Customer Review Rankings
Ralf Krestel, Nima Dokoohaki
In Elsevier Neural Networks, Volume 66, June 2015, Pages 36-45.
Abstract: E-commerce Web sites owe much of their popularity to consumer reviews accompanying product descriptions. On-line customers spend hours and hours going through heaps of textual reviews to decide which products to buy. At the same time, each popular product has thousands of user-generated reviews, making it impossible for a buyer to read everything. Current approaches to display reviews to users or recommend an individual review for a product are based on the recency or helpfulness of each review. In this paper, we present a framework to rank product reviews by optimizing the coverage of the ranking with respect to sentiment or aspects, or by summarizing all reviews with the top-K reviews in the ranking. To accomplish this, we make use of the assigned star rating for a product as an indicator for a review's sentiment polarity and compare bag-of-words (language model) with topic models (Latent Dirichlet Allocation) as a mean to represent aspects. Our evaluation on manually annotated review data from a commercial review Web site demonstrates the effectiveness of our approach, outperforming plain recency ranking by 30% and obtaining best results by combining language and topic model representations.

Pending publications

5) Saizu: Cache hit stabilization of a real dataset using a control theoretic approach.
Ian Marsh, Nima Dokoohaki, Andrea Hess, Daniel Gillblad
6) Discerning Temporal Viewing Patterns of Video-On-Demand Services Using Unsupervised Link Prediction.
Nima Dokoohaki, Henrik Abrahamsson
7) A Survey of LDA Techniques for Political Topic Distillation from Twitter.
Linn Sandberg, Nima Dokoohaki, Nina Tahmasebi

III – ATTENDED SEMINARS, WORKHOPS, CONFERENCES

- November 12, 2014 - Big Data Lab, joint activity with EU Code Week, Kista, Sweden
- November 16, 2014 - Big Data Lab, Kista, Sweden
- March 5, 2015 – National workshop on " Mobile network data for traffic applications" organized by Sweco in Stockholm, Sweden (part1)
- March 26, 2015 – National workshop on " Mobile network data for traffic applications" organized by Sweco in Stockholm, Sweden (part2)
- October 7/9, 2014 - SICS and Big Data day Kista, Sweden
- April 17, 2015- Swe-Clarin National Workshop on Digital Humanities, Göteborg University, Sweden
- (*Planned*) August 2015, attendance at TrustCom 2015
- (*Planned*) August 2015, attendance at ASONAM 2015

IV – RESEARCH EXCHANGE PROGRAMME (REP)

During the tenure of my fellowship I came in contact with Dr. Anna Monreale at Knowledge Discovery and Data Mining Lab at CNR, Pisa and Dr. Jean Marc Seigneur at Advanced Systems Group at University of Geneva. Both researchers generously showed great interest in my visit but due to unavailability of my travel documents at the time, my visits have been delayed.