REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
23-08-2019	Final Report	21-Nov-2014 - 20-May-2019

4. TITLE AND SUBTITLE
Final Report: Efficient Analytics over Hidden Online Social Networks

5a. CONTRACT NUMBER
W911NF-15-1-0020

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611104

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of Texas at Arlington
701 South Nedderman Drive
Box 19145
Arlington, TX 76019 -0145

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
66175-NS-H.20

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Gautam Das
UU	UU	UU	UU		19b. TELEPHONE NUMBER 817-272-7595

Standard Form 298 (Rev 8/98)
Prescribed by ANSI Std. Z39.18

RPPR Final Report
as of 06-Nov-2019

Agency Code:

Proposal Number: 66175NSH **Agreement Number: W911NF-15-1-0020**
INVESTIGATOR(S):

Name: Gautam Das
Email: gdas@uta.edu
Phone Number: 8172727595
Principal: Y

Organization: **University of Texas at Arlington**
Address: 701 South Nedderman Drive, Arlington, TX 760190145
Country: USA
DUNS Number: 064234610 EIN: 756000121
Report Date: 20-Aug-2019 Date Received: 23-Aug-2019
Final Report for Period Beginning 21-Nov-2014 and Ending 20-May-2019
Title: Efficient Analytics over Hidden Online Social Networks
Begin Performance Period: 21-Nov-2014 **End Performance Period:** 20-May-2019
Report Term: 0-Other
Submitted By: Gautam Das Email: gdas@uta.edu
Phone: (817) 272-7595

STEM Degrees: 2 **STEM Participants:**

Major Goals: The main objective of the project is to develop efficient analytics techniques for understanding the state of an online social network, from the organic data generated by the social network users to the behavior of these users, from the growth or decline of the social network to the user-user or user-data interactions that may contribute to such changes. The term “online social network” has a broad definition in our project, encompassing both the traditional meaning, i.e., a graph structure where nodes are users and edges represent the relationships explicitly declared by users between each other, to “virtual” networks where the nodes could be organizations or groups of users and the edges could be tacitly defined by common activities, proximity of locations, similarities of opinions, or other relationships not explicitly declared by users but implicitly inferable from data available on the online social network.

Accomplishments: Our main approach in the project is to enable the efficient analytics of online social networks, and thereby a wide range of applications from community detection to link prediction to influence maximization, etc., by addressing the following three fundamental challenges facing social-network analytics:

- Information extraction, i.e., how to transform the unstructured format of a web page displaying data from an online social network to structured formats that can then be properly cleaned, stored, and analyzed;

- Proper leverage of access interfaces, i.e., how to transform the desired analytical information such as aggregate statistics, as often defined in the form of aggregate queries, to access requests supported by the online social network, such as its web search interfaces or browsing interfaces.

- Sampling, i.e., how to identify a small number of statistically representative samples from a large-scale online social network, so as to enable fast and accurate estimations of aggregate statistics without issuing an extremely large number of access requests to the online social network.

Each of the three challenges features different scientific barriers:

- For information extraction, a key open problem is how to reach a proper balance between a manual approach, which incurs tedious labor costs but is easily customizable to each social network, and an automated approach, which can be error-prone given the fast pace and extreme flexibility of web development and implementation.

- For the proper leverage of access interfaces, a key open problem is how to understand the back-end black

boxes, such as search algorithms, ranking functions, etc., used by online social networks to process the access requests, and the implications of such black boxes on how the data retrieved from such access channels can be used for analytical purposes.

- For sampling, a key open problem is how to improve the fundamental tradeoff between bias and efficiency in the sampling process, i.e., how to reduce the bias of samples drawn from an online social network while maintaining a small cost in terms of the number of access requests issued to the social network.

We accomplished novel, practical and significant contributions on all three fronts in building a scalable analytics systems over real-world online social networks.

--Specifically, we built a human-in-the-loop system that extract structured information from the unstructured web interfaces of online social networks. This represents a significant advancement in addressing the information extraction challenge, as the system tackles the task in a semi-automated manner that requires minimal human interventions. In more detail,
we developed the HYDRA system which uses minimum human intervention to construct wrappers that automatically transform web requests and responses to structured data ready for analytics. Broadly, it consists of three major components: (1) SAMPLE-GEN which produces samples according to a given sampling distribution (2) SAMPLE-EVAL that evaluates samples produced by SAMPLE-GEN and also generates estimates for a given aggregate query and (3) TIMBR that enables fast and easy construction of a wrapper that models both input and output interface of the web database thereby translating supported search queries to HTTP requests and retrieving top-k query answers from HTTP responses.

--In terms of leveraging the access interfaces, we pioneered techniques that properly utilize a variety of access channels commonly present in real-world online social networks, e.g., k-nearest-neighbor search interface and top-k form-like search interface, to enable the efficient estimations of aggregate statistics. For the k-nearest-neighbor interface, for example, our work represent the first technique that can produce completely unbiased answers to aggregate statistics. We recognized the prevalence of location-based features in popular online social networks, which often support the search of k nearest users based on a given location. While this interface reveals highly valuable information about users, it is often severely limited in terms of the number of users returned (often 50 to 100), essentially precluding the crawl of enough users' location information to support analytical queries. To enable analytics over location information, we designed and implemented the ANALOC system, a web based platform that enables fast analytics through a k-nearest-neighbor interface by supporting the approximate processing of a wide variety of aggregate statistics over user-specified selection conditions. We also extended the ANALOC system to enable density-based clustering over real-world social networks

--Finally, we developed theoretically principled algorithms for searching the content of online social networks that provably improves the efficiency of over previous state-of-the-art techniques, and demonstrated their superiority over the existing solutions in real-world social networks. Besides the k-nearest neighbor interface, we also studied the retrieval of critical information for analytics support from a structured search interface, which is also often provided by online social networks as auxiliary means to search for contents, users, etc. Specifically, we investigated the discovery of skyline tuples (e.g., users) and the arbitrary ranking of tuples of interest.

Our research resulted in numerous publications and prototype demonstration systems that were published and/or presented in premier journals and conferences in the data management research community. The itemized list of papers appears later in this report.

Training Opportunities: This project trained multiple graduate students on conducting research related to data analytics over OSNs. Students directly supported by this project include Md. Farhadur Rahman from UTA and Yachao Lu from GWU.

RPPR Final Report
as of 06-Nov-2019

Results Dissemination: The results of this project were disseminated on premier research conferences in the field of data management. In particular, the following papers were published (and the federal support is acknowledged in them):

1. Md. Abdus Salam, M. Koone, S. B. Roy, S. Thirumuruganathan, G. Das: A Human-in-the-loop Attribute Design Framework for Classification. Accepted for publication in TheWebConf (formerly WWW) 2019.

2. Abolfazl Asudeh, Azade Nazi, Nan Zhang, Gautam Das, and H. V. Jagadish: RRR: Rank-Regret Representative. Accepted for publication in SIGMOD 2019.

3. Abolfazl Asudeh, H. V. Jagadish, Julia Stoyanovich, and Gautam Das: Designing Fair Ranking Schemes. Accepted for publication in SIGMOD 2019.

4. Abolfazl Asudeh, Azade Nazi, Nick Koudas, Gautam Das: Maximizing Gain Over Flexible Attributes In Peer to Peer Marketplaces. Accepted for publication in PAKDD 2019.

5. J. Z. Bakdash, S. Hutchinson, E. G. Zaroukian, L. R. Marusich, S. Thirumuruganathan, C. Sample, B. Hoffman, and G. Das: Malware in the Future? Forecasting of Analyst Detection of Cyber Events. Accepted for publication in the Journal of Cybersecurity, 2018.

6. Abolfazl Asudeh, Azade Nazi, Jees Augustine, Saravanan Thirumuruganathan, Nan Zhang, Gautam Das, Divesh Srivastava: Leveraging Similarity Joins for Signal Reconstruction. In PVLDB 2018.

7. Sona Hasani, Saravanan Thirumuruganathan, Abolfazl Asudeh, Nick Koudas, Gautam Das: Efficient Construction of Approximate Ad-Hoc ML models Through Materialization and Reuse. In PVLDB 2018.

8. Yeshwanth Gunasekaran, Md Farhadur Rahman, Sona Hasani, Nan Zhang, Gautam Das: DBLOC: Density Based Clustering over LOCation Based Services. Demo paper, in SIGMOD 2018.

9. Yeshwanth Gunasekaran, Abolfazl Asudeh, Sona Hasani, Nan Zhang, Ali Jaoua, Gautam Das: QR2: A Third-party Query Reranking Service Over Web Databases, Demo paper, in IEEE ICDE 2018.

10. Habibur Rahman, Senjuti Basu Roy, Saravanan Thirumuruganathan, Sihem Amer-Yahia, and Gautam Das: Optimized group formation for solving collaborative tasks, In VLDB Journal 2018.

11. Abolfazl Asudeh, Azade Nazi, Nan Zhang, Gautam Das: Efficient Computation of Regret-Ratio Minimizing Set: A Compact Maxima Representative. Proceedings of the ACM SIGMOD, 2017.

12. M. F. Rahman, W. Liu, S. Bin Suhaim, S. Thirumuruganathan, N. Zhang and G. Das: HDBSCAN: Density based Clustering over Location Based Services. Proceedings of the IEEE ICDE, 2017.

13. S. Bin Suhaim, N. Zhang, G. Das and A. Jaoua: HDBExpDetector: Aggregate Sudden-Change Detector over Dynamic Web Databases. Demo paper, IEEE ICDE, 2017.

14. Habibur Rahman, Senjuti Basu Roy, Gautam Das: A Probabilistic Framework for Estimating Pairwise Distances Through Crowdsourcing. Proceedings of EDBT, 2017.

15. A. Asudeh, N. Zhang, G. Das, Query Reranking As A Service, Proceedings of the VLDB Endowment (PVLDB), 9(11), pp. 888-899, 2016. doi: 10.14778/2983200.2983205

16. A. Asudeh, S. Thirumuruganathan, N. Zhang, G. Das, Discovering the Skyline of Web Databases, Proceedings of the VLDB Endowment (PVLDB), 9(7), pp. 600-611, 2016. doi: 10.14778/2904483.2904491

17. Y. Lu, S. Thirumuruganathan, N. Zhang, G. Das, Hidden Database Research and Analytics (HYDRA) System, IEEE Data Engineering Bulletin, 38(3), pp. 84-102, 2015. doi: 10.1145/2757302.2757306

18. Zhuojie Zhou, Nan Zhang, Gautam Das: Leveraging history for faster sampling of online social networks. In Proceedings of the VLDB Endowment, Volume 8 Issue 10, June 2015 Pages 1034-1045

19. Azade Nazi, Zhuojie Zhou, Saravanan Thirumuruganathan, Nan Zhang, Gautam Das: Walk, Not Wait: Faster Sampling Over Online Social Networks. In Proceedings of the VLDB Endowment, Volume 8 Issue 10, June 2015.

Honors and Awards: Co-PI Zhang received a Distinguished TPC Member award from the 2017 IEEE International Conference on Computer Communications (INFOCOM).

In January 2018, PI Gautam Das rank was elevated to “Distinguished University Chair” Professor of Computer Science and Engineering, University of Texas at Arlington.

Protocol Activity Status:

Technology Transfer: Nothing to Report

PARTICIPANTS:

Participant Type: PD/PI
Participant: Gautam Das
Person Months Worked: 3.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Co PD/PI
Participant: Nan Zhang
Person Months Worked: 3.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Md. Farhadur Rahman
Person Months Worked: 15.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Yachao Lu
Person Months Worked: 15.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

ARTICLES:

RPPR Final Report
as of 06-Nov-2019

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: Proceedings of the VLDB Endowment
Publication Identifier Type: DOI Publication Identifier: 10.14778/3231751.3231752
Volume: 11 Issue: 10 First Page #: 1276
Date Submitted: 10/18/18 12:00AM Date Published: 6/1/18 5:00AM
Publication Location:
Article Title: Leveraging similarity joins for signal reconstruction
Authors: Abolfazl Asudeh, Azade Nazi, Jees Augustine, Saravanan Thirumuruganathan, Nan Zhang, Gautam Da
Keywords: signal reconstruction, traffic, networks
Abstract: Signal reconstruction problem (SRP) is an important optimization problem where the objective is to identify a solution to an under- determined system of linear equations that is closest to a given prior. It has a substantial number of applications in diverse areas in- cluding network traffic engineering, medical image reconstruction, acoustics, astronomy and many more. Most common approaches for SRP do not scale to large problem sizes. In this paper, we propose a dual formulation of this problem and show how adapt- ing database techniques developed for scalable similarity joins pro- vides a significant speedup. Extensive experiments on real-world and synthetic data show that our approach produces a significant speedup of up to 20x over competing approaches.
Distribution Statement: 1-Approved for public release; distribution is unlimited.
Acknowledged Federal Support: **Y**

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: Proceedings of PAKDD 2019, Springe
Publication Identifier Type: Other Publication Identifier:
Volume: Issue: First Page #:
Date Submitted: 8/22/19 12:00AM Date Published:
Publication Location:
Article Title: Maximizing Gain Over Flexible Attributes In Peer to Peer Marketplaces.
Authors: Abolfazl Asudeh, Azade Nazi, Nick Koudas, Gautam Das
Keywords: P2P marketplaces, data mining, recom
Abstract: Peer to peer marketplaces such as AirBnB enable transactional exchange of services directly between people. In such platforms, those providing a service (hosts in AirBnB) are faced with various choices. For example in AirBnB, although some amenities in a property (attributes of the property) are fixed, others are relatively flexible and can be provided without significant effort. Providing an amenity is usually associated with a cost. Naturally different sets of amenities may have a different "gains" for a host. Consequently, given a limited budget, deciding which amenities (attributes) to offer is challenging. In this paper, we formally introduce and define the problem of Gain Maximization over Flexible Attributes (GMFA). We first prove that the problem is NP-hard and show that identifying an approximate algorithm with a constant approximate ratio is unlikely. We then provide a practically efficient exact algorithm to the GMFA problem for the general class of monotonic gain functions,
Distribution Statement: 1-Approved for public release; distribution is unlimited.
Acknowledged Federal Support: **Y**

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 1-Published
Journal: Journal of Cybersecurity
Publication Identifier Type: DOI Publication Identifier: 10.1093/cybsec/tyy007
Volume: 4 Issue: 1 First Page #:
Date Submitted: 8/23/19 12:00AM Date Published: 1/1/18 6:00AM
Publication Location:
Article Title: Malware in the future? Forecasting of analyst detection of cyber events
Authors: Jonathan Z Bakdash, Steve Hutchinson, Erin G Zaroukian, Laura R Marusich, Saravanan Thirumurugar
Keywords: malware, time series
Abstract: Cyberattacks endanger physical, economic, social, and political security. There have been exten-sive efforts in government, academia, and industry to anticipate, forecast, and mitigate such cyber-attacks. A common approach is time-series forecasting of cyberattacks based on data from networktelescopes, honeypots, and automated intrusion detection/prevention systems. This research hasuncovered key insights such as systematicity in cyberattacks. Here, we propose an alternate per-spective of this problem by performing forecasting of attacks that are “analyst-detected” and“-verified” occurrences of malware. We call these instances of malware cyber event data.Specifically, our dataset was analyst-detected incidents from a large operational ComputerSecurity Service Provider (CSSP) for the US Department of Defense, which rarely relies only onautomated systems. Our data set consists of weekly counts of cyber events over approximately7 years. This curated datas
Distribution Statement: 1-Approved for public release; distribution is unlimited.
Acknowledged Federal Support: **Y**

CONFERENCE PAPERS:

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: 2016 IEEE 32nd International Conference on Data Engineering (ICDE)
Date Received: 31-Aug-2016 Conference Date: 16-May-2016 Date Published:
Conference Location: Helsinki, Finland
Paper Title: ANALOC: Efficient analytics over Location Based Services
Authors: Md Farhadur Rahmany, Saad Bin Suhaim, Weimo Liu, Saravanan Thirumuruganathany, Nan Zhang, Ga
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: IEEE International Conference on Data Engineering (ICDE), 2017
Date Received: 09-May-2017 Conference Date: 19-Apr-2017 Date Published:
Conference Location: San Diego, CA
Paper Title: HDBExpDetector: Aggregate Sudden-Change Detector over Dynamic Web Databases
Authors: Saad Bin Suhaim, Nan Zhang, Gautam Das, Ali Jaoua
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: IEEE International Conference on Data Engineering (ICDE), 2017
Date Received: 09-May-2017 Conference Date: 19-Apr-2017 Date Published:
Conference Location: San Diego, CA
Paper Title: HDBExpDetector: Aggregate Sudden-Change Detector over Dynamic Web Databases
Authors: Saad Bin Suhaim, Nan Zhang, Gautam Das, Ali Jaoua
Acknowledged Federal Support: **Y**

RPPR Final Report
as of 06-Nov-2019

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: IEEE International Conference on Data Engineering (ICDE), 2017
Date Received: 09-May-2017 Conference Date: 19-Apr-2017 Date Published:
Conference Location: San Diego, CA
Paper Title: Density based Clustering over Location Based Services
Authors: Md Farhadur Rahman, Weimo Liu, Saad Bin Suhaim, Saravanan Thirumuruganathan, Nan Zhang, Gaut
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: ACM International Symposium on Management of Data
Date Received: 18-Oct-2018 Conference Date: 16-May-2017 Date Published:
Conference Location: Chicago, IL
Paper Title: Efficient Computation of Regret-ratio Minimizing Set: A Compact Maxima Representative
Authors: Abolfazl Asudeh, Azade Nazi, Nan Zhang, Gautam Das
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: Proceedings of the 2018 SIGMOD International Conference on Management of Data
Date Received: 18-Oct-2018 Conference Date: 10-Jun-2018 Date Published:
Conference Location: Houston, TX, USA
Paper Title: DBLOC: Density Based Clustering over LOCation Based Services
Authors: Yeshwanth Gunasekaran, Md Farhadur Rahman, Sona Hasani, Nan Zhang, Gautam Das
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: the 2017 ACM
Date Received: 18-Oct-2018 Conference Date: 06-Nov-2017 Date Published:
Conference Location: Singapore, Singapore
Paper Title: Efficient Computation of Subspace Skyline over Categorical Domains
Authors: Md. Farhad Rahman, Abolfazl Asudeh, Nick Koudas, Gautam Das
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 0-Other
Conference Name: The World Wide Web Conference
Date Received: 22-Aug-2019 Conference Date: 13-May-2019 Date Published:
Conference Location: San Francisco, CA, USA
Paper Title: A Human-in-the-loop Attribute Design Framework for Classification
Authors: Md. Abdus Salam, M. Koone, S. B. Roy, S. Thirumuruganathan, G. Das
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: the 2019 International Conference SIGMOD
Date Received: 22-Aug-2019 Conference Date: 30-Jun-2019 Date Published:
Conference Location: Amsterdam, Netherlands
Paper Title: RRR: Rank-Regret Representative
Authors: 2. Abolfazl Asudeh, Azade Nazi, Nan Zhang, Gautam Das, and H. V. Jagadish
Acknowledged Federal Support: **Y**

RPPR Final Report
as of 06-Nov-2019

Nothing to report in the uploaded pdf