Solicitation Number: W911NF-12-R-0012-01

PROPOSAL COVER PAGE

1. SUBMIT TO:
Director
U.S. Army Research Office
ATTN: AMSRL-RO-RI
P.O. Box 12211
Research Triangle Park, NC 27709-2211

2. For consideration by:
☐ Biology/Life Sci
☐ Chemistry
☒ Computer Science
☐ Electronics
☐ Mechanical
☐ Environmental Sciences
☐ Sensors & Electron Dev
☐ Vehicle Technology Dir
☐ Materials
☐ Mathematics
☐ Physics
☒ Comp & Info Sci
☐ Wepons & Mtls Sci
☐ Human Rsch & Eng
☐ Surv/Lethality

3. Is this proposal being submitted to another Federal Agency?
☑ No ☐ Yes If Yes, list the agency:

4. Is applicant delinquent on any Federal Debt?
☐ Yes (Attach explanation) ☒ No

5. Proposal Valid Until (min of 6 mos):
Januarv 1. 2014

6. Entity Identification Number (EIN) or Taxpayer Identification Number (TIN)
141368361

7. Data Universal Numbering System (DUNS No.):
804878247

8. Commercial and Government Entity (CAGE) Code: 3GPV4

9. Name of organization to which award should be made:
The Research Foundation for the State University of New York

10. Administrative Address of Organization (if different):
Office of the Vice President of Research, W5510 Frank Melville Jr. Library, Stony Brook, NY 11794-3362

11. Branch/Campus/Other Component (where work is performed, if different):
Stony Brook University

12. Submitting Organization's Contract/Grant Administration Office:
Office of the Vice President of Research

13. Submitting Organization's Audit Office:
Office of Management Analysis and Audit

14. Submitting Organization: (Check all that apply)
☐ For Profit: ☐ Large ☐ Small ☐ Disadvantaged ☐ 8a ☐ Women-Owned ☐ Foreign ☐ Individual
☒ Educational: ☐ HBCU ☐ Minority Institution ☐ Hispanic ☐ Indian Tribal ☐ State ☐ Private ☐ Foreign ☐ FDP
☐ Hospital: ☒ Public ☐ Private ☒ Nonprofit ☐ For Profit
☒ Nonprofit
☐ Not-For-Profit
☐ Other (Specify)

15. Check appropriate box(es) if this proposal includes any of the items listed below:
☐ Human Subjects ☐ Biosafety Level (BL) 1-4 Facility
☐ Vertebrate Animals ☐ Genetically Engineered Organisms
☐ National Environment Policy Act ☐ Limited Rights Data
☐ Disclosure of Lobbying Activities ☐ Unlimited Rights
☐ Historical Places ☐ Govt Purpose Rights Data
☐ GFE ☐ GFD ☐ Proprietary Data
☐ GFI ☐ GFP ☐ Ozone Depleting Substances

16. Proposed Amount:
$247,228.60

17. Proposed Duration (1-60 mos):
36 months

18. Proposed Start Date:
January 1, 2014

19. Type of Award Proposed:
☒ Single Investigator
☐ Young Investigator Program
☐ Short Term Innovation Rsch
☐ Research Instrumentation
☐ Conference/Symposia
☐ Other (Specify):

20. Title of Proposed Project: Scalable Anomaly Detection and Description in Large, Heterogeneous, Dynamic Data

21. Principal Investigator (PI)/Project Director (PD) Department and Postal Address:
Leman Akoglu, PI, Assistant Professor, Stony Brook University, Department of Computer Science, Stony Brook, NY 11794-4400

22. Year PI's degree conferred
August 2012

23. Scientific discipline of PI's degree
Ph.D., Computer Science

TYPED NAMES	TELEPHONE NUMBER	FACSIMILE NUMBER	ELECTRONIC MAIL ADDRESS
24. PI/PD Leman Akoglu	631.632.9801	631.632.8334	leman@cs.stonybrook.edu
25. CO-PI/PD			
26 a. Primary Administrative representative Authorized to Conduct Negotiations: Cornelia Seiffert, Contracts Administrator	631.632.9029	631.632.6963	Cornelia.seiffert@stonybrook.edu
26 b. Alternate Administrative Representative Authorized to Conduct Negotiations:			

27 a. Authorized Representative Signing for Applicant Organization:

27 b. Title: Department Chair

27 c. By signing and submitting this proposal, the Offeror is providing the certifications contained in this BAA.

27 d. Signature: A. Clayton Date: 05/28/2013

Form 51 (REV Jun 05)

Scalable Anomaly Detection and Description in Large, Heterogeneous, Dynamic Data

1. Introduction

The goal of this proposal is to create scalable real-time algorithms for anomaly and event *detection* in large-scale, heterogeneous, and dynamically evolving data, with additional key focus on anomaly *description*. Such dynamic and heterogeneous data frequently occur in defense, finance, media, healthcare, and communications with anomaly detection applications including detection of fraud, attacks, faults, inefficiencies, and rare events. In addition to effective and timely detection of anomalies, explaining the detected anomalies to domain analysts is a crucial part of our proposal for facilitating post-analysis and sense-making.

Specifically, our proposed research will focus on four areas:

(T1) *Real-time information-theoretic anomaly detection and description in dynamic data*, focusing on dynamic pattern-based data compression and online anomaly detection;

(T2) *Multi-resolution event detection and visualization with anomalous regions in dynamic networks*, focusing on temporal anomalies and close-by anomalous regions for visualization;

(T3) *Making or breaking temporal robustness in dynamic networks*, focusing on building algorithms to determine the best nodes/edges to add/remove to improve/degrade the robustness of a network under budget constraints and track robustness over time for event detection; and

(T4) *Anomaly detection in complex heterogeneous networks*, for spotting extremities (e.g., political propaganda, fraud) in networks with node/edge attributes and +/– edge signs.

Our research areas have direct relevance and high potential impact to the US Army; for which detecting, describing, and visualizing anomalies in complex dynamic data in real-time, and quantifying and influencing robustness of inter-connected systems constitute crucial tasks.

2. Task Descriptions

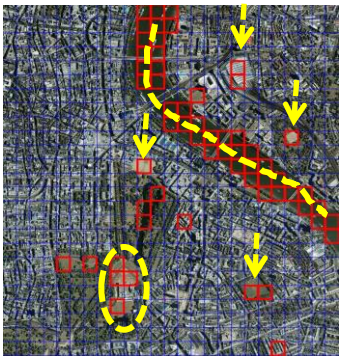
2.1 (T1) Real-time Information-Theoretic Anomaly Detection and Description in Dynamic Data

Task Description: Given a massive, high-dimensional database, how can we effectively and efficiently build a model that describes the data norms as succinctly as possible? How can we detect the database anomalies? Is there a way to explain the anomalies and how they differ from the norms? How about when the database grows over time? Answering these questions for very large databases has many key applications in finance, health care, security, law enforcement, etc.

Most work in the anomaly/outlier detection area focus on the direct detection of the anomalies in a database and completely ignore the description of the anomalies. Different from the vast majority of earlier work, our goals are to: (i) build a *compression model* that represents the *norm* of the data; (ii) not only spot but also *describe* the anomalies, and (iii) *update* the models and spot anomalies *dynamically* over time.

In our recent work [1], we developed a novel approach for anomaly detection using *pattern-based compression*. Our approach first builds a set of data compression models (in this case, code tables) to compress the database as succinctly as possible. The patterns (or code-words) in the code tables represent (information-theoretically) the best frequent patterns, i.e. the norm of the data. Based on these models, we flag those data points as anomalies that cannot be compressed well (or cannot be described well) with our models. Key advantages of our model/pattern-based approach are: (i) parameter-free nature (as it requires no user intervention),

(ii) linear computational scalability with both database size and number of dimensions, (iii) interpretability (by explaining flagged anomalies by deviations from patterns/code-words), (iv) generality (as it applies to transaction, graph, and image databases) and (v) effectiveness (see **figure** to the right, where our method detects rare regions in a given image, e.g. empty fields, shown in red tiles).



Proposed Work: In the proposed project, we plan to extend our pattern-based anomaly detection method to *dynamic* data. In our earlier work [1] we assumed that the full database is given, from which we build our compression models. On the other hand, in many real-world scenarios data keeps arriving, often at high rates, over time. Our goal will be to compress the data as it arrives and spot anomalies in real time. For such dynamic data, the main challenge is keeping our compression models up-to-date. To do so, we will update the code-word lengths of our patterns as we observe more of them, as long as the data distribution remains the same. When the underlying nature of the data changes over time (which is one of the research challenges to detect when), we will develop novel methods to merge/split and potentially discard several existing patterns as well as add new representative ones, so that the data compression is as succinct as possible. This approach will enable online, real-time anomaly detection based on the updated models. Since the models are dynamically updated and represent the data, it is not necessary to store the individual data instances, which also provides data reduction.

(T1) Problem Statement:
Given a large high-dimensional database where new data instances arrive dynamically over time;
Build a set of pattern-based compression models and **update** models as new data arrives,
Detect change points at which the data distribution changes significantly,
Spot and **describe** anomalies dynamically in real time.

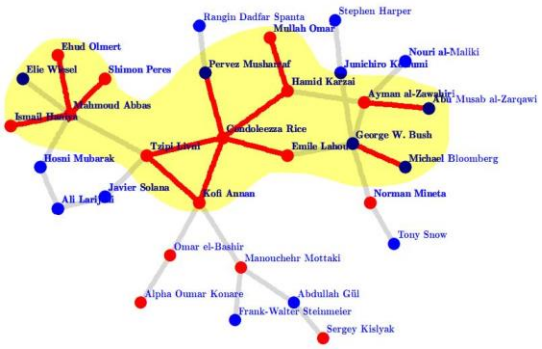
Evaluation: We will test our method on simulated network and message traffic data from our collaborator Northrop Grumman Aerospace Systems and from CAIDA (http://www.caida.org) to analyze normal and abnormal (i.e., cyber-attack) scenarios. We will also experiment with image streams (i.e., videos) to find change points and salient regions over time.
Deliverables: Novel algorithms and open-source implementation of methods for (i) dynamic data *compression* (using information-theoretic code-tables), (ii) real-time anomaly and change-point *detection*, and (iii) pattern-based anomaly *description*
Related Publication:

- [1] *Fast and Reliable Anomaly Detection in Categoric Data. Leman Akoglu, Hanghang Tong, Jilles Vreeken, and Christos Faloutsos.* ***ACM CIKM****, 2012.*

2.2 (T2) Multi-Resolution Event Detection and Visualization with Anomalous Regions in Dynamic Networks

Task Description: Given a large graph that grows and changes dynamically over time, how can we tell at what points it changes drastically, i.e. when a global event occurred? How can we detect the anomalous nodes, relations, and node groups, i.e. local anomalies, at a given time snapshot? Are those nodes/edges close-by in the graph and thus form large anomalous regions? How can we track and visualize how the anomalous regions change over time? Detecting events in relational dynamic data is a general problem setting with several real-world applications, including intrusion, fraud, and outbreak detection.

Prior research on event detection focus on finding anomalous time ticks in time series data. However, the relational nature of many applications (e.g., malicious network packages, information, diseases, etc. exchanged among agents) makes it crucial to address these problems using graph-based algorithms. While graph-based change detection, especially based on graph distance measures, has been studied in the literature, different from those works our goals are to: (i) detect events in a given graph locally at multiple resolutions, in particular at node, edge, and node groups level, (ii) identify large anomalous regions in the graph (e.g., see **figure** to the right in which red depicts anomalies and a large anomalous region is highlighted), (iii) track and visualize how the anomalous regions change over time for better sense-making.



Proposed Work: In this task, we will build novel techniques that monitor the activities in a large dynamic network over time, at multiple resolutions. Specifically, building on our prior work [2,3], we will track the time-varying behaviors of individual agents (nodes), relations/pairs of agents (edges), groups of agents (communities), and the whole network at large to flag significant changes. Importantly, given the anomalies spotted at a particular time snapshot, we will automatically determine the anomalous *regions* in the network using graph-closeness and connectivity, and also identify additional potentially-suspicious agents/relations that aid in connecting the anomalies [4]. We will exploit and expand the visualization techniques we developed [5] to show the detected anomalous regions to domain analysts, as well as how those regions change over time. We will carefully assess what to show to the analysts by focusing on the top most anomalous regions, for effective attention routing and sense-making.

(T2) Problem Statement:
Given a dynamic graph (e.g., activities over edges change, new edges/nodes appear, etc.);
Detect change-points at which the graph at large changes significantly (*global* events),
Find anomalous nodes/edges/groups of nodes that change their behavior significantly over time, (*local* events),
Find large anomalous regions of "close-by" anomalous subgraphs at *each* time tick (see figure),
Track and **visualize** anomalous regions over time.

Evaluation: We will test our method on simulated network and message traffic data from our collaborator Northrop Grumman Aerospace Systems and from CAIDA (http://www.caida.org) to analyze normal and abnormal (i.e., cyber-attack) scenarios.
Deliverables: Novel algorithms and open-source implementation of methods for (i) global and *local event detection* in dynamic graphs, (ii) finding close-by anomalous subgraphs that form large *anomalous regions* at a given time tick, and a *visualization toolkit* for (iii) visualizing to the analysts the large anomalous regions and how these regions change over time.
Related Publications:

- [2] *Event Detection in Time Series of Mobile Communication Graphs. Leman Akoglu and Christos Faloutsos.* ***27th Army Science Conference****, 2010.*
- [3] *MetricForensics: A Multi-Level Approach for Mining Volatile Graphs. Keith Henderson, Tina Eliassi-Rad, Christos Faloutsos, Leman Akoglu, Lei Li,Koji Maruhashi, B. Aditya Prakash, Hanghang Tong.****ACM SIGKDD****, 2010.*

- [4] *Mining Connection Pathways for Marked Nodes in Large Graphs. Leman Akoglu, Jilles Vreeken, Hanghang Tong, Duen Horng Chau, Nikolaj Tatti, and Christos Faloutsos.* ***SIAM SDM****, 2013.*
- [5] *TourViz: Interactive Visualization of Connection Pathways in Large Graphs. Duen Horng Chau, Leman Akoglu, Jilles Vreeken, Hanghang Tong, Christos Faloutsos.* ***ACM SIGKDD****, 2012.*

2.3 (T3) Making or Breaking Temporal Robustness in Dynamic Networks

Task Description: Robustness of a network is a measure of its resilience for targeted attacks or random failures where nodes/edges are removed from the network which might cause the network to fall apart (e.g., become disconnected). The level of connectivity of the network, e.g. the number and length of paths between the nodes, is an important factor in determining its robustness. Ideally, a fully connected network is the most robust; however it is not feasible to design fully connected real-world networks due to several constraints, such as physical space, budget, etc. Therefore, it is important to (i) quantify the robustness of a given network, (ii) identify the best way to design/attack the network to make it more/less robust under a given budget, (iii) efficiently monitor the robustness over time to spot attacks as quickly as possible.
Proposed Work: In this task, we will focus on the crucial problem quantifying and influencing the network robustness in dynamic networks.

(T3) Problem Statement:
Given a graph (and an integer budget k);
Find k edges one should add to improve the robustness of the network the most (e.g. in computer network defense) (*making* a network),
Find k edges/nodes that should be removed to destroy the robustness of a network the most (e.g. in hindering terrorist communications, or hindering disease outbreaks) (*breaking* a network),
Find the largest set of edges/nodes that can be removed from the network without affecting its robustness significantly (e.g. to sparsify the network),
Update the robustness score efficiently when the network changes over time and **spot** change-points when the robustness drops substantially.

Our goal is to build novel, fast algorithms that scale to massive networks with approximation guarantees. In our ongoing work [6], we define the robustness of a network using its natural connectivity based on its spectrum, i.e. eigen-values/eigen-vectors. To make/break the network robustness, we will create methods to identify edges/nodes the addition/removal of which changes the graph spectrum the most. We will also use techniques to quickly update the eigen-values/vectors of a graph over time to monitor the graph robustness against attacks.
Evaluation will be performed on simulated as well as real-world (dynamic) networks and algorithm performance will be measured against optimal as well as simple heuristic solutions. Real networks, e.g. DBLP, IMDB, etc., will be analyzed for how robustness changes in years.
Deliverables: Novel algorithms to quantify, make/break, monitor network robustness
Related Publication:

- [6] *Robustness under Constraints in Large Dynamic Networks. Leman Akoglu and Hanghang Tong. Work in progress.*

2.4 (T4) Anomaly Detection in Complex Heterogeneous Networks

Task Description: In our past work [7] we focused on structural anomaly detection in networks. In the real-world, the networks may be much richer, in which the nodes/edges are associated

with attributes. For example nodes in a social network may have interests or roles, relations may have characteristics such as friend/family/colleague/etc. or signs such as +/– for friend/foe relations. Building on our prior work [8], we plan to create novel data mining algorithms to spot anomalies in such heterogeneous networks. We have investigated several applications of this setting: detecting fake review(er)s [9] and detecting polarization in political forums [10] where edges have signs (+:like/for, –:dislike/against).

Proposed Work: We propose to formulate novel problem definitions of anomalies as we did in [9,10] for heterogeneous networks and create fast algorithms to extract them in large-scale graphs. Our suit of detection methods in heterogeneous settings will help spot novel anomalies beyond network structure, and facilitate the description of the anomalies to the analysts.

(T4) Problem Statement:

Given a large graph with node/edge attributes (features);
Find community anomalies, i.e. nodes with different attributes than others in the community (e.g. a Russian speaker that belongs in an English-only speaking community in a social network),
Find communities for which the distributions of (a subset of) attributes differ from majority,
Visualize network communities by both their structural *and* attribute similarities in the layout.

Evaluation: We will use social networks LiveJournal, Friendster, YouTube, and Orkut, all crawled from public data and available at http://snap.stanford.edu/data/ (links+user profiles), for qualitative evaluation. We will also inject anomalies by attribute swapping, noise addition, simulation, etc. for quantitative evaluation.

Deliverables: Anomaly detection definitions and fast algorithms for heterogeneous networks, community visualization toolkit with attribute summarization

Related Publications:

- [7] *OddBall: Spotting Anomalies in Weighted Graphs. Leman Akoglu, Mary McGlohon, Christos Faloutsos.* ***PAKDD****, June 2010.* Best Research Paper Award
- [8] *PICS: Parameter-free Identification of Cohesive Subgroups in Large Attributed Graphs. Leman Akoglu, Hanghang Tong, Brendan Meeder, Christos Faloutsos.* ***SIAM SDM****, 2012.*
- [9] *Opinion Fraud Detection in Online Reviews using Network Effects. Leman Akoglu, Rishi Chandy, and Christos Faloutsos.* ***ICWSM****, 2013. (To appear)*
- [10] *Predicting and Ranking Political Polarity in Signed Networks. Leman Akoglu. In progress.*

3. Cost Information and Research Months

Budget Estimate	**Year 1**	**Year 2**	**Year 3**
Salary - 1 GRA	$ 9,000.00	$ 9,500.00	$ 10,000.00
Tuition - 1 GRA	$ 4,188.00	$ 4,188.00	$ 4,188.00
Salary PI - 2 mos summer	$ 21,112.00	$ 21,744.00	$ 22,396.00
Fringe (PI) - 41%	$ 8,655.92	$ 8,915.04	$ 9,182.36
Fringe (GRA) - 15%..16%..17%	$ 1,350.00	$ 1,520.00	$ 1,700.00
Materials/Supplies	$ 3,000.00	$ 500.00	$ 500.00
Travel	$ 5,000.00	$ 5,000.00	$ 5,000.00
Total IDC	$ 52,305.92	$ 51,367.04	$ 52,966.36
Indirect Costs 57.5%, 58%	$ 30,075.90	$ 29,792.88	$ 30,720.49
Estimated Total Cost	**$ 82,381.82**	**$ 81,159.92**	**$ 83,686.85**
Estimated Total Project Budget			**$ 247,228.60**

Tasks	**GRA months**
(T1)	9
(T2)	12
(T3)	9
(T4)	6
Total	**36**

LEMAN AKOGLU

Assistant Professor
Department of Computer Science
Stony Brook University
Stony Brook, NY 11794

Office: +1 (631) 632-9801
Mobile: +1 (412) 996-9190
leman@cs.stonybrook.edu
http://www.cs.stonybrook.edu/∼leman

PROFESSIONAL PREPARATION

- **Carnegie Mellon University** — Pittsburgh, PA
  Ph.D., Computer Science — *Sep. 2007 - August 2012*

  Thesis Mining and Modeling Real-world Networks: Patterns, Anomalies, and Tools.
  Ph.D. advisor Christos Faloutsos. **Committee** Christos Faloutsos (Chair), Andrew Moore (CMU, Google Inc.), Aarti Singh (CMU), Andrew Tomkins (Google Inc.)

- **Bilkent University** — Ankara, Turkey
  B.S., Computer Science (with distinction) — *Sep. 2003 - May 2007*

ACADEMIC APPOINTMENTS

- **Stony Brook University** — Stony Brook, NY
  Asst. Professor (Tenure-Track), Computer Science — *August 2012 - Present*

RESEARCH INTERESTS

Keywords: Mining and modeling large-scale real-world networks, knowledge discovery, pattern mining, anomaly and event detection, graph algorithms, social network analysis, machine learning

HONORS & AWARDS

- **Received**, IBM First Patent Application Invention Achievement Award, 2012
- **Winner**, Facebook Grace Hopper Scholarship 2010
- **Winner**, Google ECML/PKDD Conference and Travel Grant Award for Women in CS, 2010
- **Winner**, Best Paper Award in PAKDD 2010
- **Winner**, Best Knowledge Discovery Paper Award in PKDD 2009
- **Winner**, "Instructors' Choice" Award for the Graphical Models class project 2008
- **Ranked 2nd**, Computer Science Department, Bilkent University, among 150 graduates, 2007
- **Ranked 300th**, National University Entrance Exam in Turkey among over 1 million students, 2003
- **Ranked 5th**, Science High School Entrance Exam in Turkey among over 300,000 students, 2000

SELECTED PUBLICATIONS

1. Leman Akoglu, Rishi Chandy, and Christos Faloutsos. *Opinion Fraud Detection in Online Reviews by Network Effects.* **ICWSM**, 2013.
2. Leman Akoglu, Jilles Vreeken, Hanghang Tong, Duen Horng Chau, Nikolaj Tatti, and Christos Faloutsos. *Mining Connection Pathways for Marked Nodes in Large Graphs.* **SDM**, 2013.
3. Leman Akoglu, Hanghang Tong, Jilles Vreeken, and Christos Faloutsos. *Fast and Reliable Anomaly Detection in Categoric Data* **ACM CIKM**, 2012.
4. Leman Akoglu, Duen Horng Chau, U Kang, Danai Koutra, Christos Faloutsos. *OPAvion: Mining and visualization in large graphs.* (Demo paper) **SIGMOD**, 2012.
5. Leman Akoglu, Hanghang Tong, Brendan Meeder, Christos Faloutsos. *PICS: Parameter-free Identification of Cohesive Subgroups in Large Attributed Graphs.* **SDM**, 2012.
6. Leman Akoglu, Mary McGlohon, Christos Faloutsos. *OddBall: Spotting Anomalies in Weighted Graphs.* **PAKDD**, 2010. **Best Research Paper**
7. Leman Akoglu and Christos Faloutsos. *RTG: A Recursive Realistic Graph Generator using Random Typing.* **ECML PKDD**, 2009. **Best Knowledge Discovery Paper**

All articles can be downloaded from `http://www.cs.stonybrook.edu/~leman/pubs.html`