List of Suggested Reviewers or Reviewers Not To Include (optional)

SUGGESTED REVIEWERS:
Not Listed

REVIEWERS NOT TO INCLUDE:
Not Listed

List of Collaborators & Other Affiliations Information

Collaborators and Co-Editors

- Salvador Aguinaga (University of Notre Dame (UND))
- George Brova (University of Illinois Urbana-Champaign (UIUC))
- Valter Crescenzi (Roma Tre)
- David Chiang (UND)
- Xiao Cheng (UIUC)
- Maria Glenski (UND)
- Thomas Gottron (Koblenz)
- Jiawei Han (UIUC)
- Thomas J. Johnston (Ebay)
- Rucha Kanade (UIUC)
- Kin Hou Lei (UIUC)
- Zuozhu Liu (Zhejiang University)
- Greg Madey (UND)
- Paul W. McBurney (University of Pennsylvania)
- Ryan McCune (Cal Poly)
- Collin McMillan (UND)
- Paolo Merialdo (Roma Tre)
- Aditya Nambiar (IIT Bombay)
- Rodrigo Palacios (California State University Fresno)
- Corey Pennycuff (UND)
- Baoxu Shi (UND)
- Greg Stoddard (Northwestern University)
- Yizhou Sun (UCLA)
- Fangbo Tao (UIUC)
- Chi Wang (Microsoft)
- Lidan Wang (Microsoft)
- Xiao Yu (Google)
- Xihao Avi Zhu (UIUC)

Graduate Advisors and Postdoctoral Sponsors

- Salvador Aguinaga (UND)
- Maria Glenski (UND)
- Ryan McCune (Cal Poly, graduated 2016)
- Corey Pennycuff (UND)
- Baoxu Shi (UND)

Thesis Advisor and Postgraduate-Scholar Sponsor

- Jiawei Han (UIUC)
- William H. Hsu (Kansas State University)

COVER SHEET FOR PROPOSAL TO THE NATIONAL SCIENCE FOUNDATION

PROGRAM ANNOUNCEMENT/SOLICITATION NO./DUE DATE NSF 15-555 07/20/16	☐ Special Exception to Deadline Date Policy	FOR NSF USE ONLY NSF PROPOSAL NUMBER
FOR CONSIDERATION BY NSF ORGANIZATION UNIT(S) (Indicate the most specific unit known, i.e. program, division, etc.) IIS - INFO INTEGRATION & INFORMATICS		

DATE RECEIVED	NUMBER OF COPIES	DIVISION ASSIGNED	FUND CODE	DUNS# (Data Universal Numbering System)	FILE LOCATION
				824910376	

EMPLOYER IDENTIFICATION NUMBER (EIN) OR TAXPAYER IDENTIFICATION NUMBER (TIN) 350868188	SHOW PREVIOUS AWARD NO. IF THIS IS ☐ A RENEWAL ☐ AN ACCOMPLISHMENT-BASED RENEWAL	IS THIS PROPOSAL BEING SUBMITTED TO ANOTHER FEDERAL AGENCY? YES ☐ NO ☒ IF YES, LIST ACRONYM(S)

NAME OF ORGANIZATION TO WHICH AWARD SHOULD BE MADE University of Notre Dame	ADDRESS OF AWARDEE ORGANIZATION, INCLUDING 9 DIGIT ZIP CODE 940 Grace Hall NOTRE DAME, IN 46556-5708
AWARDEE ORGANIZATION CODE (IF KNOWN) 0018408000	
NAME OF PRIMARY PLACE OF PERF University of Notre Dame	ADDRESS OF PRIMARY PLACE OF PERF, INCLUDING 9 DIGIT ZIP CODE University of Notre Dame Fitzpatrick Hall Notre Dame ,IN ,465565637 ,US.

IS AWARDEE ORGANIZATION (Check All That Apply) (See GPG II.C For Definitions)	☐ SMALL BUSINESS ☐ FOR-PROFIT ORGANIZATION	☐ MINORITY BUSINESS ☐ WOMAN-OWNED BUSINESS	☐ IF THIS IS A PRELIMINARY PROPOSAL THEN CHECK HERE

TITLE OF PROPOSED PROJECT **CAREER: Principled Structure Discovery for Network Analysis**

REQUESTED AMOUNT	PROPOSED DURATION (1-60 MONTHS)	REQUESTED STARTING DATE	SHOW RELATED PRELIMINARY PROPOSAL NO. IF APPLICABLE
$ 550,071	60 months	05/16/17	

THIS PROPOSAL INCLUDES ANY OF THE ITEMS LISTED BELOW

- ☐ BEGINNING INVESTIGATOR (GPG I.G.2)
- ☐ DISCLOSURE OF LOBBYING ACTIVITIES (GPG II.C.1.e)
- ☐ PROPRIETARY & PRIVILEGED INFORMATION (GPG I.D, II.C.1.d)
- ☐ HISTORIC PLACES (GPG II.C.2.j)
- ☐ VERTEBRATE ANIMALS (GPG II.D.6) IACUC App. Date ________ PHS Animal Welfare Assurance Number ________
- ☒ FUNDING MECHANISM **Research - other than RAPID or EAGER**
- ☐ HUMAN SUBJECTS (GPG II.D.7) Human Subjects Assurance Number ________ Exemption Subsection ________ or IRB App. Date ________
- ☐ INTERNATIONAL ACTIVITIES: COUNTRY/COUNTRIES INVOLVED (GPG II.C.2.j) ________
- ☒ COLLABORATIVE STATUS **Not a collaborative proposal**

PI/PD DEPARTMENT Computer Science and Engineering	PI/PD POSTAL ADDRESS 940 Grace Hall NOTRE DAME, IN 465565612 United States
PI/PD FAX NUMBER	

NAMES (TYPED)	High Degree	Yr of Degree	Telephone Number	Email Address
PI/PD NAME Tim Weninger	PhD	2013	574-631-6770	tweninge@nd.edu
CO-PI/PD				
CO-PI/PD				
CO-PI/PD				
CO-PI/PD				

CERTIFICATION PAGE

Certification for Authorized Organizational Representative (or Equivalent) or Individual Applicant

By electronically signing and submitting this proposal, the Authorized Organizational Representative (AOR) or Individual Applicant is: (1) certifying that statements made herein are true and complete to the best of his/her knowledge; and (2) agreeing to accept the obligation to comply with NSF award terms and conditions if an award is made as a result of this application. Further, the applicant is hereby providing certifications regarding conflict of interest (when applicable), drug-free workplace, debarment and suspension, lobbying activities (see below), nondiscrimination, flood hazard insurance (when applicable), responsible conduct of research, organizational support, Federal tax obligations, unpaid Federal tax liability, and criminal convictions as set forth in the NSF Proposal & Award Policies & Procedures Guide,Part I: the Grant Proposal Guide (GPG). Willful provision of false information in this application and its supporting documents or in reports required under an ensuing award is a criminal offense (U.S. Code, Title 18, Section 1001).

Certification Regarding Conflict of Interest

The AOR is required to complete certifications stating that the organization has implemented and is enforcing a written policy on conflicts of interest (COI), consistent with the provisions of AAG Chapter IV.A.; that, to the best of his/her knowledge, all financial disclosures required by the conflict of interest policy were made; and that conflicts of interest, if any, were, or prior to the organization's expenditure of any funds under the award, will be, satisfactorily managed, reduced or eliminated in accordance with the organization's conflict of interest policy. Conflicts that cannot be satisfactorily managed, reduced or eliminated and research that proceeds without the imposition of conditions or restrictions when a conflict of interest exists, must be disclosed to NSF via use of the Notifications and Requests Module in FastLane.

Drug Free Work Place Certification

By electronically signing the Certification Pages, the Authorized Organizational Representative (or equivalent), is providing the Drug Free Work Place Certification contained in Exhibit II-3 of the Grant Proposal Guide.

Debarment and Suspension Certification

(If answer "yes", please provide explanation.)

Is the organization or its principals presently debarred, suspended, proposed for debarment, declared ineligible, or voluntarily excluded from covered transactions by any Federal department or agency? Yes ☐ No ☒

By electronically signing the Certification Pages, the Authorized Organizational Representative (or equivalent) or Individual Applicant is providing the Debarment and Suspension Certification contained in Exhibit II-4 of the Grant Proposal Guide.

Certification Regarding Lobbying

This certification is required for an award of a Federal contract, grant, or cooperative agreement exceeding $100,000 and for an award of a Federal loan or a commitment providing for the United States to insure or guarantee a loan exceeding $150,000.

Certification for Contracts, Grants, Loans and Cooperative Agreements

The undersigned certifies, to the best of his or her knowledge and belief, that:
(1) No Federal appropriated funds have been paid or will be paid, by or on behalf of the undersigned, to any person for influencing or attempting to influence an officer or employee of any agency, a Member of Congress, an officer or employee of Congress, or an employee of a Member of Congress in connection with the awarding of any Federal contract, the making of any Federal grant, the making of any Federal loan, the entering into of any cooperative agreement, and the extension, continuation, renewal, amendment, or modification of any Federal contract, grant, loan, or cooperative agreement.
(2) If any funds other than Federal appropriated funds have been paid or will be paid to any person for influencing or attempting to influence an officer or employee of any agency, a Member of Congress, an officer or employee of Congress, or an employee of a Member of Congress in connection with this Federal contract, grant, loan, or cooperative agreement, the undersigned shall complete and submit Standard Form-LLL, ''Disclosure of Lobbying Activities,'' in accordance with its instructions.
(3) The undersigned shall require that the language of this certification be included in the award documents for all subawards at all tiers including subcontracts, subgrants, and contracts under grants, loans, and cooperative agreements and that all subrecipients shall certify and disclose accordingly.

This certification is a material representation of fact upon which reliance was placed when this transaction was made or entered into. Submission of this certification is a prerequisite for making or entering into this transaction imposed by section 1352, Title 31, U.S. Code. Any person who fails to file the required certification shall be subject to a civil penalty of not less than $10,000 and not more than $100,000 for each such failure.

Certification Regarding Nondiscrimination

By electronically signing the Certification Pages, the Authorized Organizational Representative (or equivalent) is providing the Certification Regarding Nondiscrimination contained in Exhibit II-6 of the Grant Proposal Guide.

Certification Regarding Flood Hazard Insurance

Two sections of the National Flood Insurance Act of 1968 (42 USC §4012a and §4106) bar Federal agencies from giving financial assistance for acquisition or construction purposes in any area identified by the Federal Emergency Management Agency (FEMA) as having special flood hazards unless the:
(1) community in which that area is located participates in the national flood insurance program; and
(2) building (and any related equipment) is covered by adequate flood insurance.

By electronically signing the Certification Pages, the Authorized Organizational Representative (or equivalent) or Individual Applicant located in FEMA-designated special flood hazard areas is certifying that adequate flood insurance has been or will be obtained in the following situations:
(1) for NSF grants for the construction of a building or facility, regardless of the dollar amount of the grant; and
(2) for other NSF grants when more than $25,000 has been budgeted in the proposal for repair, alteration or improvement (construction) of a building or facility.

Certification Regarding Responsible Conduct of Research (RCR)
(This certification is not applicable to proposals for conferences, symposia, and workshops.)

By electronically signing the Certification Pages, the Authorized Organizational Representative is certifying that, in accordance with the NSF Proposal & Award Policies & Procedures Guide, Part II, Award & Administration Guide (AAG) Chapter IV.B., the institution has a plan in place to provide appropriate training and oversight in the responsible and ethical conduct of research to undergraduates, graduate students and postdoctoral researchers who will be supported by NSF to conduct research.
The AOR shall require that the language of this certification be included in any award documents for all subawards at all tiers.

CERTIFICATION PAGE - CONTINUED

Certification Regarding Organizational Support

By electronically signing the Certification Pages, the Authorized Organizational Representative (or equivalent) is certifying that there is organizational support for the proposal as required by Section 526 of the America COMPETES Reauthorization Act of 2010. This support extends to the portion of the proposal developed to satisfy the Broader Impacts Review Criterion as well as the Intellectual Merit Review Criterion, and any additional review criteria specified in the solicitation. Organizational support will be made available, as described in the proposal, in order to address the broader impacts and intellectual merit activities to be undertaken.

Certification Regarding Federal Tax Obligations

When the proposal exceeds $5,000,000, the Authorized Organizational Representative (or equivalent) is required to complete the following certification regarding Federal tax obligations.
By electronically signing the Certification pages, the Authorized Organizational Representative is certifying that, to the best of their knowledge and belief, the proposing organization:
(1) has filed all Federal tax returns required during the three years preceding this certification;
(2) has not been convicted of a criminal offense under the Internal Revenue Code of 1986; and
(3) has not, more than 90 days prior to this certification, been notified of any unpaid Federal tax assessment for which the liability remains unsatisfied, unless the assessment is the subject of an installment agreement or offer in compromise that has been approved by the Internal Revenue Service and is not in default, or the assessment is the subject of a non-frivolous administrative or judicial proceeding.

Certification Regarding Unpaid Federal Tax Liability

When the proposing organization is a corporation, the Authorized Organizational Representative (or equivalent) is required to complete the following certification regarding Federal Tax Liability:

By electronically signing the Certification Pages, the Authorized Organizational Representative (or equivalent) is certifying that the corporation has no unpaid Federal tax liability that has been assessed, for which all judicial and administrative remedies have been exhausted or lapsed, and that is not being paid in a timely manner pursuant to an agreement with the authority responsible for collecting the tax liability.

Certification Regarding Criminal Convictions

When the proposing organization is a corporation, the Authorized Organizational Representative (or equivalent) is required to complete the following certification regarding Criminal Convictions:

By electronically signing the Certification Pages, the Authorized Organizational Representative (or equivalent) is certifying that the corporation has not been convicted of a felony criminal violation under any Federal law within the 24 months preceding the date on which the certification is signed.

Certification Dual Use Research of Concern

By electronically signing the certification pages, the Authorized Organizational Representative is certifying that the organization will be or is in compliance with all aspects of the United States Government Policy for Institutional Oversight of Life Sciences Dual Use Research of Concern.

AUTHORIZED ORGANIZATIONAL REPRESENTATIVE	SIGNATURE	DATE
NAME		

TELEPHONE NUMBER	EMAIL ADDRESS	FAX NUMBER

PROJECT SUMMARY

Overview:

This project will develop and evaluate principled techniques that learn the Lego-like building blocks of real world networks. Then, these network patterns will be used to gain insights into the mechanisms that underlie network structure and evolution.

The ideas in this proposal originate from a newfound relationship between graph theory and formal language theory discovered by the PI and his collaborators. The relationship between graph theory and formal language theory allows for a Hyperedge Replacement Grammar (HRG) to be extracted from any graph without loss of information. Like a context free grammar, but for graphs, the extracted HRG contains the precise building blocks of the network as well as the instructions by which these building blocks ought to be pieced together. Because of the principled way it is constructed, the HRG can even be used to regenerate an isomorphic copy of the original graph. By marrying the fields of graph theory and formal language theory, lessons from the previous 50 years of study in formal language theory, grammars, and much of theoretical computer science can now be applied to graph mining and network science! This proposal takes the first steps towards reconciling these disparate fields by asking incisive questions about the extraction, inference, and analysis of network patterns in a mathematically elegant and principled way.

This project will also support educational and outreach programs that will broaden participation in computer science. Open source software implementing the new algorithms will be made available to the public, and will also be made to serve as an educational tool. Research supervision and career mentoring will be made available to K-12 students through the development and publication of science fair projects in computing, and undergraduate and graduate student training will be offered through a new course in data and network science. The proposed collaborations and interdisciplinary nature of the proposed research will allow for a wide distributions of the ideas and results, which will be presented through tutorials, workshop organization, and through scholarly publications at international venues.

Intellectual Merit :

This proposal will substantially advance the state of the art in graph mining in three specific ways. (1) The PI will create precise and principled algorithms for structure discovery, extraction, sampling, and robust evaluation for static and dynamic graphs, (2) The extracted structures will be used to infer the future growth or hidden structure of the network, and (3) The discovered structures will be analyzed and mapped to real world mechanisms of network structure and growth. The proposal will also accelerate collaboration with chemical, social and natural language scientists in produce practical applications to real world data sets. Most importantly, these objectives will lay the foundation for extensive follow-up work and facilitate broad scientific impact.

Broader Impacts :

The discovery and analysis of network patterns is central to the scientific enterprise. Thus, extracting the useful and interesting building blocks of a network is critical to the advancement of many scientific fields. Indeed the most pivotal moments in the development of a scientific field are centered on discoveries about the structure of some phenomena. For example, chemists have found that many chemical interactions are the result of underlying structural properties of the interactions between elements, and biologists have agreed that tree structures are useful when organizing the evolutionary history of life. Thus, graph mining research promises to give new insights into the principles of chemistry, evolution, and ecology to name a few. Principled strategies for extracting these complex patterns are needed to discover the precise mechanisms that govern network structure and growth. This is exactly the focus of this project. Graph mining research has scientific applications of societal importance, such as new drug therapies, knowledge networks, and natural language understanding, which will be explored as part of this project. In addition, this project will result in the formation of new interdisciplinary scientists, career mentoring, graduate and undergraduate research, and the development of science fair projects in computing for K-12 students, focusing on economically disadvantaged youth.

TABLE OF CONTENTS

For font size and page formatting specifications, see GPG section II.B.2.

Appendix Items:

*Proposers may select any numbering mechanism for the proposal. The entire proposal however, must be paginated. Complete both columns only if the proposal is numbered consecutively.

CAREER: Principled Structure Discovery for Network Analysis

Tim Weninger (tweninge@nd.edu)

1 Introduction

The long term goal of the PI is to develop, study, and evaluate fundamentally new techniques for the discovery of interesting structural patterns and their function within real world networks while integrating educational and outreach programs that inspire future data mining practitioners and researchers. The goal of this CAREER proposal is to develop and evaluate principled techniques that learn the lego-like building blocks of real world networks. Then, these network patterns will be used to gain insights into the mechanisms that underlie network structure and evolution.

The ideas in this proposal originate from a newfound relationship between graph theory and formal language theory discovered by the PI and his collaborators. The relationship between graph theory and formal language theory allows for a *Hyperedge Replacement Grammar* (HRG) to be extracted from any graph without loss of information. Like a context free grammar, but for graphs, the extracted HRG contains the precise building blocks of the network as well as the instructions by which these building blocks ought to be pieced together. Because of the principled way it is constructed, the HRG can even be used to regenerate an isomorphic copy of the original graph.[5] By marrying the fields of graph theory and formal language theory, lessons from the previous 50 years of study in formal language theory, grammars, and much of theoretical computer science can now be applied to graph mining and network science! This proposal takes the first steps towards reconciling these disparate fields by asking incisive questions about the extraction, inference, and analysis of network patterns in a mathematically elegant and principled way.

The discovery and analysis of network patterns is central to the scientific enterprise. Thus, extracting the useful and interesting building blocks of a network is critical to the advancement of many scientific fields. Indeed the most pivotal moments in the development of a scientific field are centered around discoveries about the structure of some phenomena.[63] For example, chemists have found that many chemical interactions are the result of underlying structural properties of interactions between elements.[25,29] Biologists have agreed that tree structures are useful when organizing the evolutionary history of life,[30,56] sociologists find that triadic closure underlies community development,[32,44] and neuroscientists have found "small world" dynamics within neurons in the brain.[12,15] In other instances, the structural organization of the entities may resemble a ring, a clique, a star, a constellation, or any number of complex configurations. Unfortunately, current graph mining research deals with small pre-defined patterns[57,77] or frequently reoccurring patterns,[27,48,55,64–66] even though interesting and useful information may be hidden in unknown and non-frequent patterns. Principled strategies for extracting these complex patterns are needed to discover the precise mechanisms that govern network structure and growth. This is exactly the focus of this project: **to develop and evaluate techniques that learn the building blocks of real world networks that, in aggregate, succinctly describe the observed interactions expressed in the network**.

The key insight for this task, described in detail in Section 2.3, is that a network's *clique tree* (also known as the tree decomposition, junction tree, intersection tree, or cluster graph, depending on the context) encodes robust and precise information about the network. An HRG, which is extracted from the clique tree, contains graphical rewriting rules that can match and replace graph fragments similar to how a Context Free Grammar (CFG) rewrites characters in a string. These graph fragments represent a succinct, yet complete description of the building blocks of the network, and the rewriting rules of the HRG represent the instructions on how the graph is pieced together. Although the isomorphic guarantees are exciting and important, this proposal will focus instead on finding meaning in the building blocks and their instructions.

This proposal will substantially advance the state-of-the-art in graph mining in three specific ways: (1) The PI will create precise and principled algorithms for structure discovery, extraction, sampling, and robust evaluation for static and dynamic graphs, (2) The extracted structures will be used to infer the future growth and hidden structure of the network, and (3) The discovered structures will be analyzed and mapped to real world mechanisms of network structure and growth. This proposal will also accelerate collaboration with chemical, social, and natural language scientists to produce practical applications of principled structure discovery on real world data.

Education and Outreach In parallel to these research objectives, this proposal will launch educational initiatives that promote critical thinking, boost recruitment, and broaden the appeal of data science to new audiences. To that end, the PI will dedicate time and resources to teaching and mentoring undergraduate and graduate students while also enhancing K-12 curricula through a collaboration with the Northern Indiana Regional Science and Engineering Fair (NIRSEF) as well as local public school math and science teachers. Data mining and network science are ideal fields for developing STEM researchers due to their inherently interdisciplinary nature and for being some of the fastest growing and in-demand fields worldwide. The educational vision is of a large, diverse, and talented pool of scientists presenting cutting edge research where: (1) Primary and secondary school children experiment with and present science fair projects in computing, and (2) a robust and meaningful relationship is crated between high school students from under-represented groups and university scholars that will motivate and support both students and researchers throughout their careers.

1.1 Context and Long Term Objectives

The PI's past and current research experiences are well suited to successfully perform the proposed work. Graduate training at the University of Illinois Urbana-Champaign, sponsored by the NSF GRFP and NDSEG fellowships, provided a strong background in data mining and network science. At the intersection of these fields, the PI has contributed to the development of data mining tools that are capable of uncovering interesting and useful patterns that are hidden in graphical data.[49–51,58,124,125] For example, research during graduate school resulted in algorithms that uncovered hidden hierarchical structures comprising knowledge graphs like Wikipedia, citation networks and most Web sites.[75,111,113–115,117] As an Assistant Professor at the University of Notre Dame, the PI continues to apply this philosophy while working with the Air Force Office of Scientific Research on patterns of social media behavior (FA9550-15-1-0003, project ends July 2017),[110,116,119] and with the Templeton Foundation on network structures in knowledge graphs (FP053369-M/O, project ended May 2016).[4,94–97] Although similar in their underlying principles, the past projects do not overlap with the objectives of this proposal.

2 Background and Related Work

2.1 Graph Mining

For the purposes of the current proposal, graph mining technologies can be divided into two classes: (1) subgraph mining algorithms and (2) graph generating models.

Subgraph Mining Rooted in data mining and knowledge discovery, subgraph mining methods are efficient and scalable algorithms for traditional frequent itemset mining on graphs.[43,54] Frequent graph patterns are subgraphs that are found from a single large graph or a collection of many smaller graphs. A subgraph is deemed to be frequent if it appears more than some user-specified support threshold. Being descriptive models, frequent subgraphs are useful in characterizing graphs and can be used for clustering, classification or other discriminative tasks. Because of their nature, these methods have a so-called "combinatorial explosion" problem[106] wherein the search space grows exponentially with the pattern size. This causes computational headaches, but also returns a massive result set that hinders real world applicability. Recent work that heuristically mines graphs for important or representative subgraphs have been developed in response, but are still limited by their choice of heuristic.[76,84,103,120] Alternatively, researchers characterize a network by counting small subgraphs called graphlets, and therefore forfeit any chance of finding larger, more interesting structures.[8,78,86] Overcoming these limitations will require a principled approach that discovers the structures within graphs and is the first research objective of the proposed work.

Graph Generators Graph generators, like frequent subgraph mining, find distinguishing characteristics of networks, but go one step further by generating new graphs that "look like" the original graph(s). What a graph looks like includes local graph properties like the counts of frequent subgraphs described above, but can also include global graph properties like the degree distribution, clustering coefficient, diameter, and assortativity among many others. Early graph generators, like the random graph of Erdős and Reyni,[33] the small world network of Watts and Strogatz,[109] or the scale free graph of Albert and Barabási,[11] did not learn a model from a graph directly, but rather had parameters that could be tuned to generate graphs with certain desirable properties. Recent work in exponential random graphs,[90] Kronecker graphs,[17,71] Chung-Lu graphs,[23] and their many derivatives[61,82,83,85] create a model from some example graph in order to generate a new graph that has many of the same *global* properties as the original graph.

Despite their conceptual similarity, subgraph mining algorithms and graph generators have little in common algorithmically. Simply put, they solve different problems. Although Kronecker and Chung-Lu graph generators learn their model from an exemplar graph, only the exponential random graph (ERG) model actually learns a model based on the specific patterns and structures found in the graph. Unfortunately, the ERG model must pre-define the space of possible structures, and the complexity of the ERG model is exponential in the number and size of the pre-defined structures. The standard ERG model is not good at generating new graphs; however the learned model can be informative about the nature of the underlying graph, albeit through the lens of only a handful of small structures, *e.g.*, edges, triangles, 4-cliques.[41] The proposed work will bridge the gap between subgraph mining and graph generation to create a new suite of models and tools that can not only create informative models of real world data, but also generate, extrapolate, and infer new graphs in a precise, principled way.

2.2 Clique trees and Hyperedge Replacement Grammars

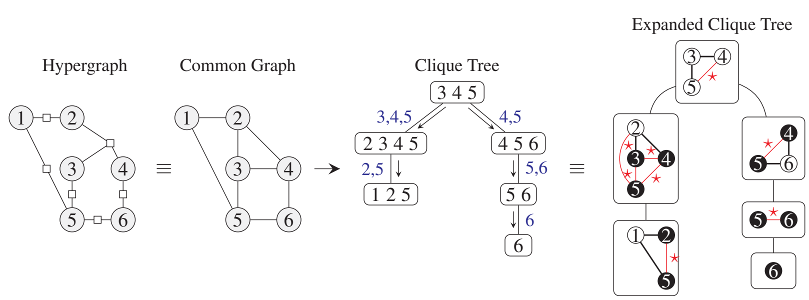


Figure 1: Example hypergraph and common graph representation (on left). A clique tree (middle) is constructed from an elimination ordering over the graph. An expanded view of the clique tree (on right) shows constituent subgraphs with triangulated edges labeled with a ⋆.

In graph theory, all graphs can be decomposed (though not uniquely) into a *clique tree*.[62] A clique tree of any graph (or any hypergraph) is a tree, each of whose nodes is labeled with nodes and edges from the original graph, such that *vertex cover*, *edge cover* and the *running intersection* properties hold,[89] and the "width" of the optimal clique tree measures how tree-like a graph is. The reason for the wide interest in finding the clique tree of a graph is because many computationally difficult problems can be solved efficiently when the data is constrained to be a tree; so by decomposing the graph into a clique tree certain problems can be solved efficiently. Figure 1 illustrates the clique tree of a graph, and how the expanded clique tree represents patterns in the network.

Within data mining and machine learning, clique trees are best known for their role in exact inference in probabilistic graphical models, constraint satisfaction, and query optimization. Unfortunately, finding the optimal, *i.e.*, the minimal-width, clique tree is NP-Complete.[9] However many reasonable approximations exist for general graphs[13,105] and the discovery of better algorithms is an active area of research in discrete

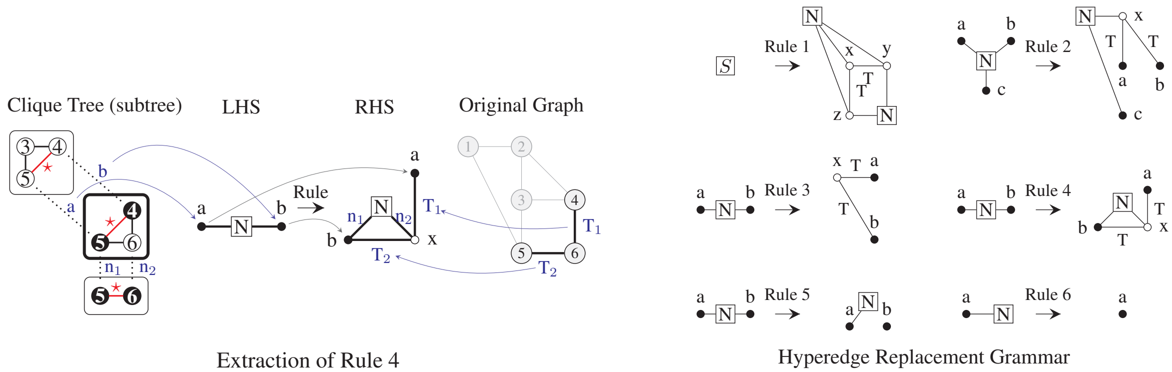


Figure 2: Example extraction of an HRG rule from the clique tree (on left). Full HRG with six rules extracted from the clique tree (on right). ○ denotes new nodes created by a RHS labeled with x, y, z, etc., ● denotes existing nodes that match earlier rules via the sepset labeled with a, b, etc.

mathematics.[3,74,80]

HRGs are a graphical counterpart to context free string grammars used in compilers and natural language processing.[31] Like in a context free string grammar, an HRG contains a set of production rules $\mathcal{P}$, each of which contains a left hand size (LHS) A and a right hand size (RHS) R. In context free string grammars, the LHS must be a nonterminal character, which can be replaced by some set of nonterminal or terminal characters on the RHS of the rule. In HRGs, nonterminals are graph-cliques and a RHS can be any graph (or hypergraph) fragment.

Just as a context free string grammar generates a string, an HRG can generate a graph by repeatedly choosing a nonterminal A and rewriting it using a production rule $A \to R$. The replacement hypergraph fragment R can itself have other nonterminal hyperedges, so this process is repeated until there are no more nonterminals in the graph.

Clique trees and HRGs have been studied separately for some time in discrete mathematics and graph theory literature. HRGs are conventionally used to generate graphs with very specific structures, *e.g.*, rings, trees, stars. A drawback of many current applications of HRGs is that their production rules must be manually defined. For example, the production rules that generate a ring-graph are distinct from those that generate a tree, and defining even simple grammars by hand is difficult or impossible. Very recently, Kemp and Tenenbaum developed an inference algorithm that learned probabilities from real world graphs, but they still relied on a handful of rather basic hand-drawn production rules (of a related formalism called vertex replacement grammar) to which probabilities were learned.[57] Kukluk, Holder and Cook were able to define a grammar from frequent subgraphs,[27,48,64–66] but their methods have a coarse resolution because *frequent* subgraphs only account for a small portion of the overall graph topology.

2.3 The Missing Link

The work proposed in this CAREER proposal is based, in part, on the relationship between a clique tree of a graph and HRGs. This relationship was first introduced theoretically by Lautemann in 1988,[69] but did not have an algorithmic solution until Gildea found a limited algorithmic solution in 2011 for use in grammar parsing[39] that was adapted by Chiang *et al.* to parse natural language.[22] Very recently, the PI developed a general solution to this challenge that allows an HRG to be extracted from any graph or hypergraph in a principled way, and that HRG can be used to generate an isomorphic copy of the original graph.[5]

The details of the algorithmic solution are straightforward, but they are not presented in this proposal. Instead, a small example of the extraction process is illustrated on the left side in Fig. 2. Here a production rule is created from the perspective of a node in the clique tree from Fig. 1. The LHS is a *nonterminal* hyperedge with a size equal to the size of the *sepset* between the current node and its parent; the RHS contains *nonterminal* hyperedges corresponding to the sepset between the current node and its children, and *terminal* hyperedges that are copied from the original graph. Because they correspond to the sepset of a child-node, nonterminal edges can be further replaced by rules created further down the clique tree. Leaf nodes in the clique tree must therefore produce a RHS with only terminal edges. This process will produce one rule for every clique tree node in a top-down manner; *e.g.*, the six nodes in the clique tree produce the six rules illustrated on the right of Fig. 2.

Unlike existing approaches, an ordered application of the rules in the extracted grammar will produce an isomorphic copy of the original graph, even if the clique tree is non-optimal. To see that this is the case, start with Rule 1 from the HRG in Fig. 2 and apply each rule in order. The result, shown in Fig. 4 on Page 8 will be isomorphic to the original graph. Unfortunately, keeping the proper rule ordering takes as much space as the original graph. However, the extracted rules can also be applied stochastically to generate, extrapolate, or otherwise create new graphs that share properties that are similar to the original graph.

These are very exciting results. This newfound ability to extract a grammar from a graph has the ability to merge two large fields of computer science. Network scientists and graph mining researchers can use principles discovered by formal language researchers and theorists; and formal language researchers may be able to apply graphical principles to their work.

Work has just begun. There are several avenues that bare the potential for groundbreaking scholarship. As a graph mining researcher, the PI will lead a research effort that not only solves important challenges, but also proposes new ones and forms partnerships to explore and understand these topics.

3 Research Plan

The research goal of this CAREER proposal is to study, develop, characterize, and evaluate techniques that use HRGs to discover and understand the structure and growth of real world networks in a principled way. To support this goal, the following objectives will be accomplished:

Objective 1: Precise and complete structure discovery, including extraction, sampling, and robust evaluation protocols will be developed and vetted.

Objective 2: Principled graph generation will be demonstrated and studied using the discovered structures on static and evolving data.

Objective 3: An analysis of the discovered structures and their relationships to real world phenomena will be theoretically studied and experimentally evaluated.

3.1 Project Overview

To achieve precise and complete structure discovery of a real world network, two essential requirements must be met within a single system:

i. The building blocks, *i.e.*, small subgraphs, that comprise any real world network must be efficiently and exactly captured in the model, and
ii. The model must represent the local and global patterns that reside in the data.

The first requirement overcomes limitations that are found in state-of-the-art graph mining algorithms. By extracting an HRG from the graph's clique tree the model will capture all the necessary graph building blocks. The second requirement is met by the relationship between graphs and context free grammars. Because of the significance of the proposed work, there is an enormous amount of research that needs to be done. Among the many possibilities, the three objectives detailed in this CAREER proposal were chosen because they have the most potential for broad impact and open the door for the widest followup work.

It is important to note that the success of each task is not dependent on the success of any other task. Many tasks could be executed in parallel, but individual progress will be measured and evaluated separately,

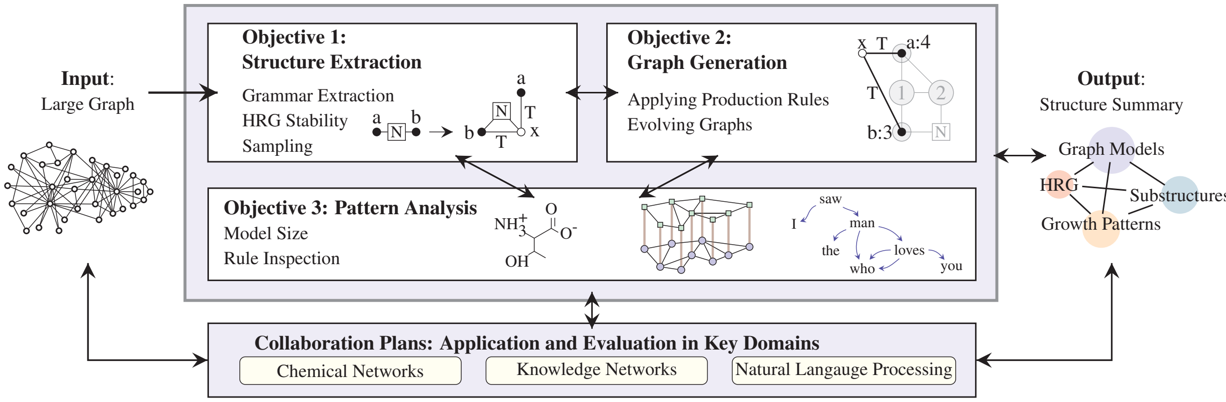


Figure 3: How research tasks interact and benefit from each other.

and then integrated into a single system. Following current lab practices, system design and implementation will be open source and will incorporate continuous feedback from collaborators.

This project does not involve **human or animal subjects**. However, the PI does have experience with IRB protocol and will receive prior approval if necessary.

Objective 1: Network Structure Extraction

A hyperedge replacement grammar (HRG) is able to represent any graph (or hypergraph) structure, but not uniquely. That is, a single graph can be represented by many different clique trees, and even an optimal clique tree may not be unique. Production rules are directly extracted from the clique tree, so it is important to understand how the choice of tree decomposition algorithm and the shape of the clique tree affects the grammar.

Task 1.1: Model Stability

Finding an optimal tree decomposition and corresponding minimal-width clique tree is NP-Complete.[9, 122] Fortunately, many reasonable approximations exist for general graphs. The PI's preliminary work employed the commonly used maximum cardinality search (MCS) algorithm introduced by Tarjan and Yannikakis[105] in 1985. MCS is a straightforward algorithm that creates a reasonable, but probably non-optimal, clique tree. A surge in recent theoretical and application-oriented projects has made a tremendous impact by finding bounded and near-optimal heuristics for real-world graphs.[3, 14, 18, 20] Each tree decomposition algorithm has certain heuristics and implementation decisions that are unavoidable; these decisions may introduce bias, which may affect the shape of the clique tree. For example, the MCS algorithm chooses an *elimination ordering*, *i.e.*, the ordering of nodes in the clique tree, based, in part, on the number of edges each node has.

Because the HRG is directly extracted from the clique tree, the choice of tree decomposition algorithm raises several questions: (1) How much does the choice of tree decomposition algorithm affect the shape of the clique tree? (2) How similar (or dissimilar) are the clique trees produced by multiple runs of a tree decomposition algorithm? and (3) How *stable* is the extracted HRG given various clique trees, *i.e.*, do different clique tree produce the same or different HRGs?

A goal of this task is to understand the relationship between a tree decomposition algorithm, its clique tree, and the extracted HRG. The choice of tree decomposition algorithm, shape of the clique tree, and the extracted HRG will be rigorously investigated using tree distance metrics, and standard statistical analysis. Even though the resulting clique trees may prove to be of different shapes, the extracted HRG may still be stable because the node labels are not copied into the grammar. It is therefore possible, even likely, that different clique trees will still produce very similar HRGs.

No matter the outcome, further interesting questions can be asked and answered. Because of the HRG-to-graph relationship, if the extracted HRGs are indeed vastly different, then the production rules that do overlap will be uniquely informative about the nature of the data. If the extracted HRGs are similar, then the extracted HRG will be uniquely representative of the data.

Evaluation Plan The PI possesses hundreds of graph datasets originating from public online resources like SNAP, KONECT and the UCI Repository, as well as several graph decomposition algorithms that can be used in experimental tests.[3,13,14,74,122] To answer questions about stability, a principled notion of tree and grammar similarity is required. Many principled metrics exist for tree similarity,[121] but the PI will need to make some adaptations to account for items unique to a clique tree like the sepset. Comparing grammars is one area where results from formal language theory may be helpful. Unfortunately, the problem of determining whether two different grammars produce the same string is undecidable,[47] which means that exact comparison between HRGs is undecidable. Nevertheless, just as in formal language theory, many approximate similarity methods exist for CFGs that can be adapted for HRGs.[68]

Here the central theme of this CAREER proposal is evident: **the PI will be able to adapt and leverage ideas and approaches from computational and formal language theory to solve difficult challenges in graph mining and network analysis**.

Task 1.2: Subgraph Sampling

Despite their relative efficiency, near-optimal tree decomposition algorithms can still be impractical on very large graphs. Rather than computing the clique tree and HRG from a graph in its entirety, it is possible to create many small HRGs from subgraph samples of the original graph and merge the sampled-HRGs into a single, representative HRG. Of course, sampling almost certainly eliminates the possibility of regenerating an isomorphic copy of the original graph, but it opens the door for many interesting questions: (1) What would it mean if the HRGs from various subgraph samples are similar? What if they are different? (2) How should small HRGs be merged to create a large, representative HRG? and (3) Does a sampled-and-merged HRG look similar to an HRG created from a full, non-sampled clique tree?

The goal of this task is to answer these questions by experimenting with different subgraph sampling and merging regimens. Subgraph sampling is an active area of research from which several principled algorithms can be drawn.[7,52,72,93] Many existing subgraph sampling algorithms try to find samples that share properties of the global network, only on a smaller scale. Other algorithms look for representative samples that find subgraphs that have a wide range of properties. Naïve sampling methods simply perform a random walk or breadth first search starting at random nodes.[107] The choice of sampling algorithm is sure to affect the structures present in the extracted HRG, and the findings from the HRG Stability Task (1.1) will provide a robust understanding of structure extraction in general.

Merging several HRGs created from subgraph samples of the same graph is another important decision that needs to be understood. The naïve way is to perform a type of set-union operator to create a single, large, hopefully representative HRG. A more principled approach can again be found by leveraging the methods and theorems from formal language theory, where model merging (and splitting) of CFGs has been heavily investigated.[67,102]

Evaluation Plan Different subgraph sampling algorithms will be evaluated by comparing subgraph HRGs to each other, and by comparing the merged HRG to an HRG created from a non-sampled graph. Ideally, the merged HRG will resemble the non-sampled HRG, and the HRG comparison method created in Task 1.1 will be employed here to quantify the similarity. Higher similarity is better. In cases where the full HRG cannot be computed, due to massive size or other constraints, the PI will use downstream tasks such as those described in Objectives 2 and 3 to further evaluate the sampled grammar.

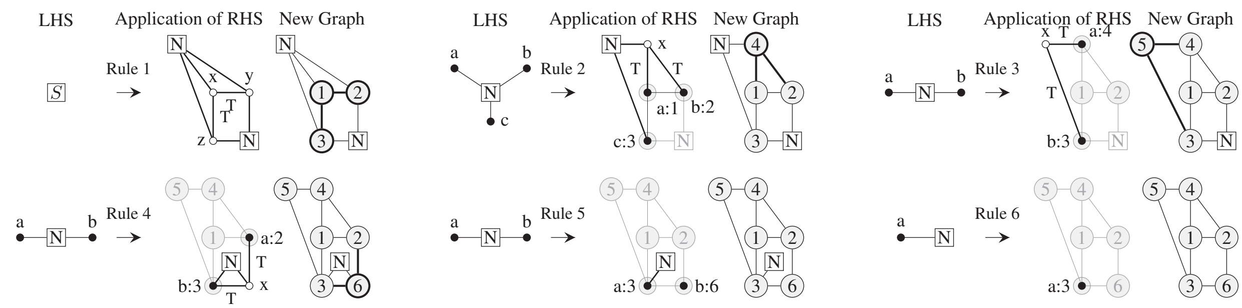


Figure 4: Ordered application of rules from the HRG in Fig. 2 creates an isomorphic copy of the original graph.

Risk Management Graph sampling techniques are well understood within the graph mining community, but finding a principled way to merge HRGs may prove to be a difficult task. The notion of *substitutability* will be especially helpful for a proper merge operation,[24,28] but the key will be to find an algorithmic solution that is efficient on HRGs. If needed, the set-union operator will suffice for downstream tasks, although the model size would probably have redundant rules and patterns.

Objective 2: Network Inference

The second objective will explore network inference, that is, the generation, extrapolation, and prediction of networks based on an example graph. This is an important objective because many networks are incomplete or are subsets of the whole data. In other cases, privacy concerns or ethical obligations prohibit the release of the whole or exact network. For example, although Facebook is unwilling to make their entire social network public, they may be willing to release an HRG, which can be used to closely, although not exactly, represent the social network. Finally, many networks are not done growing. By understanding the exact patterns of network development, a temporal HRG may be able to predict future network growth.

Task 2.1: Applying Production Rules to Generate Static Graphs

An isomorphic copy of the original graph can only be reconstructed if a traversal ordering over the clique tree is kept. Figure 4 illustrates this rewriting process using a perfect ordering. This is not practical or even desired under many conditions, so it is important to understand when and how production rules are picked when running a simulation or building a new graph. A random application of rules is unlikely to be helpful because the user will have no control over how big the network will grow. An unlucky draw of mostly terminal rules, *i.e.*, rules that do not have any nonterminal edges like Rule #3 or Rule #6 from the example HRG in Fig. 2, will result in a very small graph. Conversely, an unlucky drawing of nonterminal rules may cause a graph to grow uncontrollably to a massive size. This raises critical questions: (1) How should rules be chosen in order to maximize the similarity of the generated graph with the original graph? and (2) Can an HRG extracted from a graph sample (or many samples) be used to generate a new graph that is similar to the whole, original graph?

To answer the first question the PI will, yet again, leverage recent discoveries in formal language theory. In this case, an exact-size graph can be generated by adapting algorithms on *probabilistic* context free grammars (PCFGs),[19,59] which is currently used to generate exact-length sentences in natural language processing tasks[1,21,37,100,101] and has also been adapted for generating nucleic acid sequences.[60] Rather than randomly or stochastically drawing production rules, a graphical adaptation of PCFGs will be able to create a graph with a user-specified number of nodes or edges. Furthermore, because of their close relationship to Markov models, graphical PCFGs will be able to model the generative process in limited, but informative ways. For example, a graphical PCFG may re-discover processes, like Yule's preferential attachment,[126] Watts' random reattachment,[108,109] or some other undiscovered process, in a sound, principled manner.

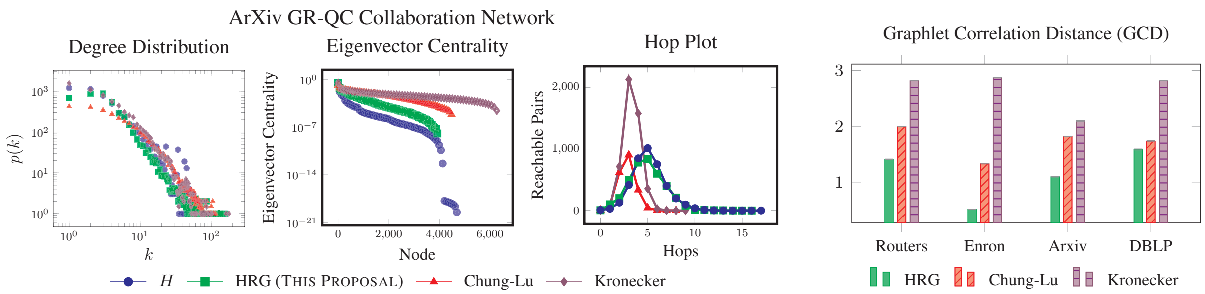


Figure 5: Example hypergraph H, common graph representation and minimal clique tree CT constructed from an optimal elimination ordering over H. Edge labels represent the intersection, *i.e.*, sepset, of the cliques in CT.

Evaluation Plan Production rule applications will be compared to other state-of-the-art graph generators including the exponential random graph model,[90] the Chung Lu model[23] (and its derivatives), the Kronecker model,[71,73] and other graph generators. Using a broad corpus of real world networks retrieved from online repositories such as SNAP or KONECT as well as those collected through the PI's collaborators (see page 12). Evaluation will be performed by comparing many global and local graph properties including the degree distribution, the principle eigenvector, clustering coefficient, diameter, hop plot, and many others.

These network properties primarily focus on statistics of the global network. However, there is mounting evidence that argues that a comparing two networks' graphlet distributions is a better way to measure the network similarity. Graphlets succinctly describe the number of small, local substructures that compose the overall graph and therefore more completely represent the details of what a graph "looks like".[86] Recent work from systems biology has identified a new metric called the Graphlet Correlation Distance (GCD) that measures the distance between the graphlet distributions of two graphs.[123] As an example, the plots in Fig. 5 show preliminary results comparing ArXiV's GR-QC collaboration network H against graphs generated via a random drawing of rules from an extracted HRG, as well as graphs generated with the Chung-Lu and Kronecker models.

The PI will further test subgraph sampling and grammar robustness using similar methodology, and with special attention paid to the local structures of the generated graph. *Graph extrapolation* will be evaluated by extracting an HRG from a subgraph of the original graph, generating a new graph of the full size, and comparing the original graph to the generated graph. *Grammar robustness* will be evaluated by recursively extracting a grammar from a generated graph. Recent work by the PI argues that this recursive generation test could also be used as a performance metric for graph generator robustness in general.[6]

Task 2.2: Applying Production Rules to Grow Dynamic Graphs

It is important to remember that a grammar extracted from a static graph can only speculate on the processes that *may* have generated the graph. The actual growth patterns can only be modelled in the presence of temporal data, where the addition or deletion of nodes and edges are shown in chronological order or marked with a timestamp. The addition of a temporal component raises interesting questions: (1) How can HRGs be adapted to model evolving graphs? (2) Can growth patterns be extracted from the grammar? and (3) Can temporal HRGs be used to predict the future topology of a graph?

To answer these questions, the PI will investigate temporal HRG extraction. Because evolving graphs add their nodes and edges a few at a time, there will be no need to perform a tree decomposition to find a clique tree. Instead, a change in the graph is analogous to the traversal of an edge in a clique tree: see Fig. 6 for an example. Just as with sepsets in the clique tree, the new LHS will need to be added to a RHS in some previous production rule. Due to the *running intersection property*, graph updates may need to propagate through much of the grammar. This could be a laborious task. The addition of a single edge is easy to extract, but not very informative; temporal binning of the network updates will also need to be explored in order to extract reasonable production rules. Even with evolving networks, there may be lessons to learn from formal language theory where CFGs have been used in timeseries analysis.[81,91]

With a temporally extracted HRG the answer to the second question will become more clear. An ordered application of the HRG rules will result in an isomorphic copy of the original graph. But, most importantly, the order of production rules will show how and when certain important structures were formed, *e.g.*, how a wedge became a triangle, how two nodes crept topologically closer and closer together before finally connecting, or how any number of complex configurations came into being. By adding a short memory to a Markov chain of production rules, graph inference will be performed by applying the graph's own growth patterns. The PI will use this information to extract new and interesting temporal patterns that govern the generative processes that build real world networks.

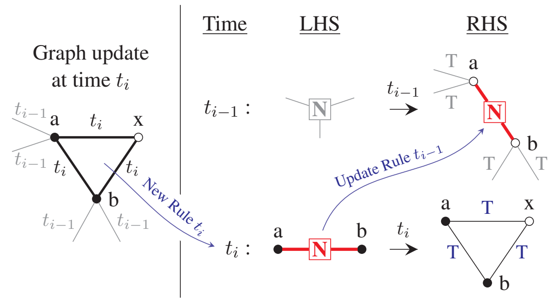


Figure 6: On left: the addition of a triangle between one new node (labelled x) and two existing nodes (labelled a and b) at timestep t_i. This change is reflected by the addition of rule t_i with a size-2 nonterminal on the LHS and a triangle on the RHS; the added nonterminal must be propagated to earlier rules.

Evaluation Plan An evaluation of the proposed approaches will answer the third question: can temporal HRGs predict the future of a graph? By understanding the patterns and processes that generated a graph to time t, the PI will be able to probabilistically re-apply production rules to generate predictions of the future evolution of the graph. Validation will be performed by hiding the last k timesteps, generating an HRG, and predicting the final k timesteps. Precision and recall metrics will be adapted to measure the performance.

Risk Management It is easy to naively generate graphs from HRGs, *c.f.*, Fig. 4. The adaption of string grammar algorithms on PCFGs will require effort, but is a low risk task. In dynamic graphs, propagating new rules through the grammar is an algorithmically expensive task. If this task proves too expensive, it is possible to extract rules by working backwards in time, but this may require significant memory. If needed, a forward-backward hybrid approach will be created to fit computational and memory bounds.

Objective 3: Pattern Analysis

The number of rules, even unique rules, that are extracted from a single graph can be quite large. Establishing a tradeoff between model size and accuracy via some sparsity condition or sampling procedure may be warranted in certain conditions. Furthermore, the "important" rules in a reduced model are likely to reveal interesting properties about real world phenomena that were previously hidden. The potentially large number of HRG rules poses several interesting challenges: (1) What are the effects of subgraph sampling on model size and performance? (2) Can the number of HRG rules be reduced in a principled way without negatively affecting model performance? (3) Can the production rules be mined to identify existing and new real world phenomena? and (4) What does it mean if two graphs have similar grammars?

Task 3.1: Model Size

As new rules are extracted from a graph they may or may not match an existing rule in the HRG. Rather than keeping a list of rules, the HRG can be treated as a multiset of rules with a counter of the number of times a particular rule has been seen. In this case, the growth of an HRG will almost certainly follow Heaps law of diminishing returns. Despite this sublinear growth, large graphs may still result in large models.

The goal of this task is to reduce the size of the HRG without negatively effecting the performance of downstream tasks like network inference. The subgraph sampling task (1.2) is a first step in reducing the model's size. Preliminary results, illustrated in Fig. 7 on the ArXiV GR-QC collaboration network, show that larger subgraph samples do result in more accurate network generation. Fortunately, a close inspection of these results shows diminishing returns, *i.e.*, an increase in model size does not correspond to an equal increase in performance.

In addition to subgraph sampling, a second objective is to reduce the number of HRG rules, *i.e.*, the model size, by reformulating the network generation task as a sparse coding task. In general terms, sparse coding is a class of unsupervised learning algorithms that learn a handful of important *basis vectors* from some data using a *loss function* and a *regularizer*.[70] In terms of HRGs and graph representation, the loss function will be defined as the difference between a generated graph and the original graph, and a regularizer will be created so as to minimize the number of basis vectors, *i.e.*, production rules, in the model.

Unfortunately, the search space is probably not convex, so an optimally sparse set of production rules may not be easily obtainable. In that case, the PI can still use any number of learning algorithms, especially inference engines used throughout natural language processing, to find reasonable results.[26,88]

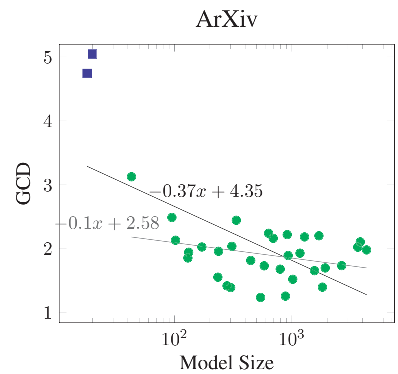


Figure 7: Model performance (lower is better) as a function of model size.

Evaluation Plan Model size will be evaluated by counting the rules. The PI will also adapt the Akaike and Bayesian information criterion (AIC/BIC) to measure the quality of the models. As in Objective 2, model performance will be measured by comparing generated graphs against the original graphs using various graph properties.

Task 3.2: Rule Inspection

HRG production rules represent the building blocks of a network in an interpretable way. Therefore, a principled investigation of common rules or reoccurring rule-combinations is likely to result in interesting new discoveries from various network datasets. For example, the triadic closure process is known by social scientists to underlie community development, cooperative behavior, and trust to name a few.[2,10,36,79] Of course, triadic closure is a known process, *i.e.*, scientists already know what to look for. Uncovering new, interesting, unknown processes is more difficult. Fortunately, the learning process in the previous task (3.1), simplifies this objective by providing a sparse HRG model, where only the important rules are identified.

It is likely that two different graphs covering the same type of data, *e.g.*, two different collaboration networks or two different social networks, will have similar HRGs. But what if they do not? Understanding the difference between two or more grammars may be an important indicator in understanding certain properties of each graph. Furthermore, if two graphs from different fields of science result in similar HRGs, then their similarity may be indicative of some broader phenomena. The technology proposed in this proposal will be employed to answer these questions and provide unprecedented opportunities to discover the hidden building blocks of network data.

Evaluation Plan Rule inspection can be evaluated using any number of standard data mining metrics, but the true test of these methods and models will be in the quantity and quality of natural phenomenon that are captured and understood through collaboration with subject matter experts.

Risk Management The rule inspection task, and many of the collaborative efforts depend on learning the important network patterns and corresponding HRG rules. This may be a difficult task. If needed, the sparse coding task can be replaced with SVMs, graph-based Principle Component Analysis, or many other statistical models. If need be, the rule probabilities assigned via the PCFG in Task 2.1 can be used to pick the most important rules.

4 Incisive Network Analysis through Collaborative Partnerships

The PI is a member of Notre Dame's Interdisciplinary Center for Network Science and Applications (iCeNSA), a collaboration of physical, social, and computer scientists with the overarching goal of facilitating and accelerating partnerships between biologists, chemists, physicists, sociologists, etc. with computer scientists, especially in the area of network science and data mining. This CAREER proposal will accelerate these partnerships for long term impact. Three such partnerships are described here.

Chemical Networks Transformations and interactions between molecules have innumerable impacts on industry and human health. The bonds of molecules are themselves a graph, and the transformations between molecules performed naturally and artificially are large and complex graphs. Prof. John Parkhill, Assistant Professor of Analytical Chemistry (see letter) at UND develops faster and more accurate approaches to calculate the energies of molecules, which enables the accurate prediction of connections between molecules without an intractable number of real experiments. Theoretical features of these graphs will yield new information about the way scientists explore chemical diversity, and the underlying design principles of chemical networks, and may support developments in the discovery of therapeutic molecules. The Parkhill group will generate a large library of molecular transformations and use HRGs to study these chemical networks.

Knowledge Networks The explosion of digital information offers an unprecedented opportunity to study the dynamics that shape human understanding, investigation, and certainty. By applying HRG extraction and mining to information networks, the PI will be able to better understand how humans create and organize the artifacts of knowledge. Prof. James Evans, Professor of Sociology at the University of Chicago (see letter), will provide expertise in the dynamic social processes by which humans create and consume knowledge networks. Prof. Evans' MetaKnowledge Network, of which the PI is an active member, contains scientists from dozens of universities across nearly all scientific fields and is currently constructing an extremely large knowledge network from scientific literature.[35,38,92,98] This ongoing collaboration has yielded four peer-reviewed publications on the Wikipedia knowledge network (and its derivatives, *e.g.*, DBpedia) for fact checking, human navigation,[4] and link prediction.[95–97] New work, supported by this CAREER proposal, will extract network patterns from the full *Web of Science* dataset, recently curated by the Metaknowledge group. This highly sought after dataset presents an unprecedented opportunity to explore the dynamics of knowledge networks.

Natural Language Processing Because HRGs are based on formal language theory, natural language processing (NLP) is an obvious application area. Recent developments in NLP have found that abstract meaning representations (AMR) are able to represent the meaning of large and complex sentences by encoding the meaning of a sentence as a graph.[45,87] Graphical representations of natural language enables the PI to discover new and interesting patterns of meaning that frequently occur in natural language. Prof. David Chiang, Associate Professor of Computer Science at UND (see letter), recently showed that AMRs could be parsed from natural language.[22] HRGs can be used to represent sentence-sized AMRs, but larger linguistic structures representing complex thoughts or stories can also be represented as a single, large AMR-network. In this form, Prof. Chiang will work closely with the PI to discover common linguistic patterns within a story or work of literature. HRGs extracted from AMR-graphs could also be compared across disciplines or cultures to provide a unique, principled understanding of human communication.

5 Education and Outreach Plan

The PI is committed to education and outreach objectives that increase and broaden computer science instruction in primary and secondary schools. Despite appeals from the White House and Congress to strengthen pre-college computer science education, the unfortunate reality is that many primary and high school students do not have access to computer science or programming curricula. In cases where computer science AP courses are available only 21% of the enrollees are female, and only 8% are African-American or Latino students.[34] This needs to change. The future of our discipline relies heavily on the curiosity of the next generation, and so it is critical that we instill in them a passion for learning and discovery. To help address these challenges, the PI will:

1. Better prepare computer science students for complex, emerging problems by: a) integrating scientific discoveries from across disciplines, and b) developing interdisciplinary research skills through new curricula, supplements and educational programs listed in Section 5.1.
2. Develop future computer scientists by integrating computing projects into widely accessible K-12 science fair projects and through unique outreach activies as described in Section 5.2.

5.1 Integrating Research into Curriculum Development Activities

The PI will train undergraduate and graduate students from computing, engineering, and natural sciences through the development of interdisciplinary courses. For example, the PI has previously developed and taught a course since 2013 titled "Web Science and Information Retrieval." Despite being primarily a graduate-level course, to date 78% of all students have been undergraduate students, and about one-third of students major in a field other than computer science. Student evaluations have rated this course in the top 25% of computer science courses for the past two years. The PI will integrate research into the classroom by creating a new course on network science and graph mining; this course will not only serve to recruit students into the PI's research group, but also to train a new generation of interdisciplinary scientists in network science and data mining using a common terminology and a unifying point of view.

Finding meaningful links between domains has been shown to be a strong driver of discovery.[16,42,53] In existing computer science curricula, graph theory and network science is typically reserved as a senior elective or purely graduate level course, while theory of computation is often a sophomore or junior level course. To bridge this gap and to promote this research, the PI will create and post online a text and video supplement to the standard undergraduate literature on context free grammars. Interestingly, Sipser's seminal textbook on "Theory of Computation" introduces formal languages and grammars immediately after its introduction of graphs.[99] So, in order to reach the widest possible audience the PI's supplement on graph grammars will match Sipser's notation and be written to an undergraduate audience.

Student-Centered Pedagogical Strategies. The PI also teaches the core undergraduate course on "Database Concepts." In both courses the PI will continue to implement an active learning classroom style, which punctuates 15-minute chalk-and-talk lectures with two minute group exercises. Two minute group exercises, especially, change the pedagogical structure by allowing students who understand the exercise to effectively instruct their neighbor in a one-on-one setting, which increases the understanding of all students. The 15-minute mini-lectures will also be designed with an eye towards a future online adaptation. Term projects will be designed to motivate students and encourage practical creativity; to date, at least six (of about 40 total) course projects have matured into downloadable apps or have otherwise been integrated into long-term solutions, *e.g.*, Notre Dame's Nutritional Accounting System, St. Ed's Diner Ordering System.

Interdisciplinary Research and Educational Programs. The PI is closely involved in several ongoing education, training, and outreach initiatives that deepen and broaden scientific impact. The PI regularly advises and often publishes with undergraduate students from NSF REU projects, Notre Dame's International Summer Undergraduate Research Experience (iSURE) program, and has mentored several internal undergraduate projects. The PI is also part of a large NSF Research Experience for Teachers (RET) program (NSF-1609394) that aims to engage high school teachers in state-of-the-art data mining and machine learning research. The PI has organized tutorials and workshops at world-class venues such as WWW, KDD, WSDM; is a speaker and panelist at SIGKDD's broadening participation in data mining (BPDM) initiative; and has presented many invited talks including a TEDx talk that was selected as the TED "talk of the month."

Forming a Strong Research Program One of the PI's long term goals is to mentor graduate students to be independent and ethical scholars with strong communication skills. The PI's lab is currently composed of four students from Hispanic, American, and Asian backgrounds; three male and one female. A fifth co-advised student, whom predates the PI's arrival at Notre Dame, has graduated and will begin a tenure-track professorship at Cal Poly this Fall. These graduate students have won several prestigious awards and fellowships including the IBM PhD Fellowship, the USAID Fellowship, and the NSF EAPSI Fellowship as well as numerous university-level awards. In addition, the PI has mentored eight undergraduate research projects resulting in four publications in international venues[4, 117–119] and numerous local posters and presentations.

5.2 Science Fair Projects in Computing

For most Americans, a science fair is the first (and perhaps most memorable) exposure to the scientific process, and recent initiatives from Google and the White House to promote science fair participation has had a measurable impact on science fair participation. However, projects in computing are largely absent from school science fairs. With this in mind, the PI has been actively involved in the Northern Indiana Regional Science and Engineering Fair (NIRSEF) for the past three years. Although national statistics are not available, NIRSEF had 293 participants in 2016 but only **four** computer science projects! This discrepancy is caused by a confluence of factors including the documented lack of K-12 CS curriculum and educators; but the PI has also found a dearth of science fair resources available to aspiring scientists.

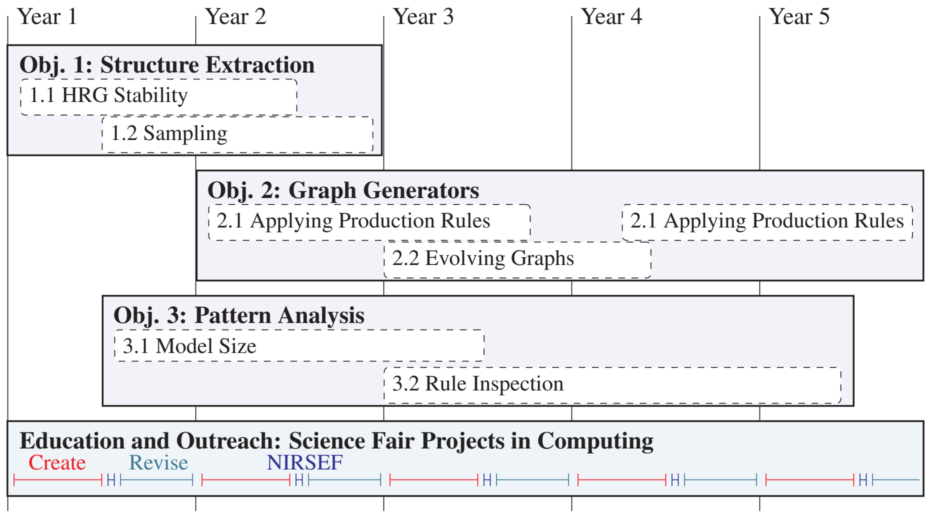
Development and Dissemination of CS Science Fair projects. Many science fair projects are adaptations of online resources: including hundreds of chemistry, biology, physics, mechanical, and electronics project ideas. Yet there are few, if any, computing projects listed online. To remedy this discrepancy, the PI, in collaboration with NIRSEF (see letter) and several local K-12 science teachers, will develop and disseminate interesting and innovative computing science fair projects.

These projects will be grouped into 7-9 and 10-12 grade levels, where the highest group will be subdivided into junior and senior-level projects. It is not reasonable to expect many K-12 students to have standard laptops or PCs, especially in South Bend, IN, a city where 65.1% of public school students qualify for free or reduced lunches. Because mobile devices are increasingly being used as the primary or sole computer, many of the science fair projects will be designed to be conducted on the students' mobile phone or tablet. Projects like: 'how the number of open apps affects battery life,' or 'mapping wireless and wifi signal in the school,' among many others, will inspire deeper questions in battery technology and wireless networks just as model volcanoes inspire deeper questions in chemistry and geology.

Community Outreach to Middle and High schools. This proposal will also initiate a novel collaboration with Mr. George Logdson, a math and science teacher at Riley High School in South Bend (see letter). The PI will extend Mr. Logdson's involvement in the aforementioned RET program to advertise and mentor computing science fair projects, and to collect feedback from students in order to create new projects and revise existing projects. The PI will also work with Notre Dame's Kaneb Center for Teaching and Learning (see letter) to develop evaluation measures that properly gauge student learning outcomes. An undergraduate student is budgeted to help with the science fair and evaluation tasks. Projects that are successful on a local level will be hosted on NIRSEF's resources Web page for world-wide dissemination. Although letters are not included, the PI has similar agreements with four other local math and science teachers through NIRSEF and the RET program.

6 Broader Impacts of Proposed Research and Educational Activities

The impact of the education and outreach initiatives will increase the number and diversity of graph mining, data mining and data science researchers. In addition, this project will result in the formation of new interdisciplinary scientists, career mentoring, undergraduate and graduate research, and the development of computer science projects for K-12 students, focusing on economically disadvantaged youth.



Work Plan: Illustration of the timeline of the proposed tasks. The success of each individual task will impact, but not preclude the success of downstream tasks. Educational tasks alternate between project development, presentation at NIRSEF, and subsequent revision. The outreach tasks are inherently cyclic and are centered around the annual science fair, which takes place in early March.

The research goal of this CAREER proposal is to develop and evaluate innovative techniques that learn the building blocks of real world networks and the instructions by which the pieces fit together. This component has the potential to dramatically enable groundbreaking advances in graph mining and networks science and directly benefit society through an improved understanding of the fundamental building blocks of real world networks. This CAREER proposal will be the first to leverage the relationship between graph theory and formal language theory. Stemming from this CAREER proposal, many of the principled lessons, theorems, and algorithms that have been developed in formal language theory will be applied to graph theory. The three objectives will not only answer important questions, but also lay the foundation for extensive followup work and facilitate broad scientific impact.

Open Source Science The PI has a history of releasing research artifacts including source code and data. All research papers will be published to ArXiV upon acceptance. Because of the broad appeal of the objectives in this CAREER proposal, the developed models and techniques will be integrated into an open source tool and released under the liberal CC-BY license. The software will be configured for easy adaptation by network scientists, and easy use by domain scientists via a user-friendly and platform independent user interface as a result of the collaborative efforts. The software will also be used as an educational tool through an online Web application. Because of this commitment to open source science, the PI's previous work has been implemented or adapted by others at IBM, Google, Facebook, Reddit, and in several large academic demonstrations.[40,46,97,104,112,117]

7 Summary

To recap, the PI's long term goal is to study new techniques for the discovery of structure in real world networks while developing educational and outreach initiatives that broaden, motivate, and inspire future computer scientists. The research objectives in this CAREER proposal leverage the PI's unique background and recent work to fundamentally change the way we view structure discovery in networks, and computer and natural scientists alike will have powerful new tools by which to understand the organization and evolution of real world networks.

Results from Prior NSF Support

The PI has not received prior NSF support.

References Cited

[1] S. Abney, D. McAllester, and F. Pereira. Relating probabilistic grammars and automata. In *ACL*, pages 542–549. Association for Computational Linguistics, 1999.

[2] L. A. Adamic and E. Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003.

[3] A. B. Adcock, B. D. Sullivan, and M. W. Mahoney. Tree decompositions and social graphs. *Internet Mathematics*, 2016.

[4] S. Aguinaga, A. Nambiar, Z. Liu, and T. Weninger. Concept hierarchies and human navigation. In *BigData*, pages 38–45. IEEE, 2015.

[5] S. Aguinaga, R. Palacios, D. Chiang, and T. Weninger. Growing graphs from hyperedge replacement grammars. In *CIKM*. ACM, 2016.

[6] S. Aguinaga and T. Weninger. The infinity mirror test for analyzing the robustness of graph generators. In *KDD Workshop on Mining and Learning with Graphs*. ACM, 2016.

[7] N. K. Ahmed, J. Neville, and R. Kompella. Network sampling: From static to streaming graphs. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 8(2):7, 2014.

[8] N. K. Ahmed, J. Neville, R. A. Rossi, and N. Duffield. Efficient graphlet counting for large networks. In *ICDM*, pages 1–10. IEEE, 2015.

[9] S. Arnborg, D. G. Corneil, and A. Proskurowski. Complexity of finding embeddings in ak-tree. *SIAM Journal on Algebraic Discrete Methods*, 8(2):277–284, 1987.

[10] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *SIGKDD*, pages 44–54. ACM, 2006.

[11] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[12] D. S. Bassett and E. Bullmore. Small-world brain networks. *The neuroscientist*, 12(6):512–523, 2006.

[13] H. L. Bodlaender. A linear-time algorithm for finding tree-decompositions of small treewidth. *SIAM Journal on computing*, 25(6):1305–1317, 1996.

[14] H. L. Bodlaender and A. M. Koster. Treewidth computations i. upper bounds. *Information and Computation*, 208(3):259–275, 2010.

[15] E. Bullmore and O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, 2009.

[16] S. Carey. *Conceptual change in childhood*. MIT press, 1985.

[17] D. Chakrabarti, Y. Zhan, and C. Faloutsos. R-mat: A recursive model for graph mining. In *SDM*, volume 4, pages 442–446. SIAM, 2004.

[18] F. Chang, C.-Y. Guo, X.-R. Lin, and C.-J. Lu. Tree decomposition for large-scale svm problems. *Journal of Machine Learning Research*, 11(Oct):2935–2972, 2010.

[19] E. Charniak, S. Goldwater, and M. Johnson. Edge-based best-first chart parsing. In *Proceedings of the sixth workshop on very large corpora*, pages 127–133. Citeseer, 1998.

[20] C. Chekuri and J. Chuzhoy. Large-treewidth graph decompositions and applications. In *STOC*, pages 291–300. ACM, 2013.

[21] Z. Chi. Statistical properties of probabilistic context-free grammars. *Computational Linguistics*, 25(1):131–160, 1999.

[22] D. Chiang, J. Andreas, D. Bauer, K. M. Hermann, B. Jones, and K. Knight. Parsing graphs with hyperedge replacement grammars. In *ACL*, pages 924–932, 2013.

[23] F. Chung and L. Lu. The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences*, 99(25):15879–15882, 2002.

[24] A. Clark. Learning trees from strings: a strong learning algorithm for some context-free grammars. *Journal of Machine Learning Research*, 14(1):3537–3559, 2013.

[25] B. L. Clarke. Theorems on chemical network stability. *The Journal of Chemical Physics*, 62(3):773–775, 1975.

[26] S. B. Cohen and M. Collins. A provably correct learning algorithm for latent-variable pcfgs. In *ACL*, pages 1052–1061, 2014.

[27] D. J. Cook and L. B. Holder. Substructure discovery using minimum description length and background knowledge. *Journal of Artificial Intelligence Research*, 1:231–255, 1994.

[28] F. Coste, G. Garet, and J. Nicolas. A bottom-up efficient algorithm learning substitutable languages from positive examples. In *ICGI*, volume 34, pages 49–63, 2014.

[29] G. Craciun and C. Pantea. Identifiability of chemical reaction networks. *Journal of Mathematical Chemistry*, 44(1):244–259, 2008.

[30] W. F. Doolittle and E. Bapteste. Pattern pluralism and the tree of life hypothesis. *Proceedings of the National Academy of Sciences*, 104(7):2043–2049, 2007.

[31] F. Drewes, H.-J. Kreowski, and A. Habel. Hyperedge replacement, graph grammars. *Handbook of Graph Grammars*, 1:95–162, 1997.

[32] D. Easley and J. Kleinberg. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press, 2010.

[33] P. Erdos and A. Rényi. On the evolution of random graphs. *Bull. Inst. Internat. Statist*, 38(4):343–347, 1961.

[34] B. Ericson. Detailed ap cs 2013 results: Unfortunately, much the same, January 2014.

[35] J. A. Evans and P. Aceves. Machine translation: Mining text for social theory. *Annual Review of Sociology*, 42(18), 2016.

[36] S. L. Feld. The focused organization of social ties. *American journal of sociology*, pages 1015–1035, 1981.

[37] S. Geman and M. Johnson. Dynamic programming for parsing and estimation of stochastic unification-based grammars. In *ACL*, pages 279–286. Association for Computational Linguistics, 2002.

[38] A. Gerow, B. Lou, E. Duede, and J. Evans. Proposing ties in a dense hypergraph of academics. In *International Conference on Social Informatics*, pages 209–226. Springer International Publishing, 2015.

[39] D. Gildea. Grammar factorization by tree decomposition. *Computational Linguistics*, 37(1):231–248, 2011.

[40] M. Glenski and T. Weninger. Rating effects on social news posts and comments. *Trans. on Intelligent Systems and Technology*, 7(5), 2016.

[41] A. Goldenberg, A. X. Zheng, S. E. Fienberg, and E. M. Airoldi. A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2(2):129–233, 2010.

[42] A. Gopnik, A. N. Meltzoff, and P. Bryant. *Words, thoughts, and theories*, volume 1. Mit Press, 1997.

[43] G. Grahne and J. Zhu. Fast algorithms for frequent itemset mining using fp-trees. *TKDE*, 17(10):1347–1362, 2005.

[44] M. S. Granovetter. The strength of weak ties. *American journal of sociology*, pages 1360–1380, 1973.

[45] J. Groschwitz, A. Koller, and C. Teichmann. Graph parsing with s-graph grammars. In *ACL*, pages 1481–1490, 2015.

[46] J. Hailpern, N. D. Venkata, and M. Danilevsky. Truncation: all the news that fits we'll print. In *ACM symposium on Document engineering*, pages 165–174. ACM, 2014.

[47] J. Hartmanis. Context-free languages and turing machine computations. In *Proceedings of Symposia in Applied Mathematics*, volume 19, pages 42–51, 1967.

[48] L. B. Holder, D. J. Cook, and S. Djoko. Substucture discovery in the subdue system. In *KDD workshop*, pages 169–180, 1994.

[49] W. H. Hsu, A. L. King, M. S. Paradesi, T. Pydimarri, and T. Weninger. Collaborative and structural recommendation of friends using weblog-based social network analysis. *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 6:55–60, 2006.

[50] W. H. Hsu, J. Lancaster, M. S. Paradesi, and T. Weninger. Structural link analysis from user profiles and friends networks: A feature construction approach. In *ICWSM*. AAAI, 2007.

[51] J. Huang, H. Sun, Y. Liu, Q. Song, and T. Weninger. Towards online multiresolution community detection in large-scale networks. *PloS one*, 6(8):e23829, 2011.

[52] C. Hübler, H.-P. Kriegel, K. Borgwardt, and Z. Ghahramani. Metropolis algorithms for representative subgraph sampling. In *ICDM*, pages 283–292. IEEE, 2008.

[53] B. Inhelder. *The early growth of logic in the child: Classification and seriation*, volume 83. Routledge, 2013.

[54] C. Jiang, F. Coenen, and M. Zito. A survey of frequent subgraph mining algorithms. *The Knowledge Engineering Review*, 28(01):75–105, 2013.

[55] I. Jonyer. Graph grammar learning. *Mining Graph Data*, pages 183–201, 2006.

[56] S. A. Kauffman. *The origins of order: Self organization and selection in evolution*. Oxford University Press, USA, 1993.

[57] C. Kemp and J. B. Tenenbaum. The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105(31):10687–10692, 2008.

[58] S. Kim, H. Kim, T. Weninger, J. Han, and H. D. Kim. Authorship classification: a discriminative syntactic tree mining approach. In *SIGIR*, pages 455–464. ACM, 2011.

[59] D. Klein and C. D. Manning. Accurate unlexicalized parsing. In *ACL*, pages 423–430. Association for Computational Linguistics, 2003.

[60] B. Knudsen and J. Hein. Rna secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, 15(6):446–454, 1999.

[61] T. G. Kolda, A. Pinar, T. Plantenga, and C. Seshadhri. A scalable generative graph model with community structure. *SIAM Journal on Scientific Computing*, 36(5):C424–C452, 2014.

[62] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[63] T. S. Kuhn. *The structure of scientific revolutions*. University of Chicago press, 2012.

[64] J. Kukluk, L. Holder, and D. Cook. Inferring graph grammars by detecting overlap in frequent subgraphs. *International Journal of Applied Mathematics and Computer Science*, 18(2):241–250, 2008.

[65] J. P. Kukluk, L. B. Holder, and D. J. Cook. Inference of node replacement recursive graph grammars. In *SDM*, pages 544–548. SIAM, 2006.

[66] J. P. Kukluk, C. H. You, L. B. Holder, and D. J. Cook. Learning node replacement graph grammars in metabolic pathways. In *BIOCOMP*, pages 44–50, 2007.

[67] K. Kurihara and T. Sato. Variational bayesian grammar induction for natural language. In *International Colloquium on Grammatical Inference*, pages 84–96. Springer, 2006.

[68] K. Lari and S. J. Young. The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer speech & language*, 4(1):35–56, 1990.

[69] C. Lautemann. Decomposition trees: structured graph representation and efficient algorithms. In *CAAP'88*, pages 28–39. Springer, 1988.

[70] H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. In *Advances in neural information processing systems*, pages 801–808, 2006.

[71] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani. Kronecker graphs: An approach to modeling networks. *Journal of Machine Learning Research*, 11(Feb):985–1042, 2010.

[72] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *SIGKDD*, pages 631–636. ACM, 2006.

[73] J. Leskovec and C. Faloutsos. Scalable modeling of real graphs using kronecker multiplication. In *ICML*, pages 497–504. ACM, 2007.

[74] B. Li, F. Z. Moataz, N. Nisse, and K. Suchan. Minimum size tree-decompositions. *Electronic Notes in Discrete Mathematics*, 50:21–27, 2015.

[75] C. X. Lin, B. Zhao, T. Weninger, J. Han, and B. Liu. Entity relation discovery from web tables and links. In *WWW*, pages 1145–1146. ACM, 2010.

[76] W. Lin, X. Xiao, and G. Ghinita. Large-scale frequent subgraph mining in mapreduce. In *ICDE*, pages 844–855. IEEE, 2014.

[77] M. H. Luerssen. Graph grammar encoding and evolution of automata networks. In *Australasian conference on Computer Science-Volume 38*, pages 229–238. Australian Computer Society, Inc., 2005.

[78] D. Marcus and Y. Shavitt. Rage–a rapid graphlet enumerator for large networks. *Computer Networks*, 56(2):810–819, 2012.

[79] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.

[80] O. J. Mengshoel. Understanding the scalability of bayesian network inference using clique tree growth curves. *Artificial Intelligence*, 174(12):984–1006, 2010.

[81] D. Minnen, I. Essa, and T. Starner. Expectation grammars: Leveraging high-level expectations for activity recognition. In *CVPR*, volume 2, pages II–626. IEEE, 2003.

[82] S. Mussmann, J. Moore, J. J. Pfeiffer, and J. Neville III. Assortativity in chung lu random graph models. In *Workshop on Social Network Mining and Analysis*, page 3. ACM, 2014.

[83] S. Mussmann, J. Moore, J. J. Pfeiffer III, and J. Neville. Incorporating assortativity and degree dependence into scalable network models. In *AAAI*, pages 238–246, 2015.

[84] S. Nijssen and J. N. Kok. The gaston tool for frequent subgraph mining. *Electronic Notes in Theoretical Computer Science*, 127(1):77–87, 2005.

[85] J. J. Pfeiffer, T. La Fond, S. Moreno, and J. Neville. Fast generation of large scale social networks while incorporating transitive closures. In *SocialCom Workshop on Privacy, Security, Risk and Trust (PASSAT)*, pages 154–165. IEEE, 2012.

[86] N. Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, 2007.

[87] M. Pust, U. Hermjakob, K. Knight, D. Marcu, and J. May. Parsing english into abstract meaning representation using syntax-based machine translation. *Training*, 10:218–021, 2015.

[88] A. Ratnaparkhi. A maximum entropy model for part-of-speech tagging. In *EMNLP*, volume 1, pages 133–142. Philadelphia, USA, 1996.

[89] N. Robertson and P. D. Seymour. Graph minors. ii. algorithmic aspects of tree-width. *Journal of algorithms*, 7(3):309–322, 1986.

[90] G. Robins, P. Pattison, Y. Kalish, and D. Lusher. An introduction to exponential random graph (p*) models for social networks. *Social networks*, 29(2):173–191, 2007.

[91] M. S. Ryoo and J. K. Aggarwal. Recognition of composite human activities through context-free grammar based representation. In *CVPR*, volume 2, pages 1709–1718. IEEE, 2006.

[92] A. Rzhetsky, J. G. Foster, I. T. Foster, and J. A. Evans. Choosing experiments to accelerate collective discovery. *Proceedings of the National Academy of Sciences*, 112(47):14569–14574, 2015.

[93] L. Schietgat, F. Costa, J. Ramon, and L. De Raedt. Effective feature construction by maximum common subgraph sampling. *Machine Learning*, 83(2):137–161, 2011.

[94] B. Shi and T. Weninger. Mining interesting meta-paths from complex heterogeneous information networks. In *ICDM Workshops*, pages 488–495. IEEE, 2014.

[95] B. Shi and T. Weninger. Discriminative predicate path mining for fact checking in knowledge graphs. *Knowledge-Based Systems*, 104:123–133, 2016.

[96] B. Shi and T. Weninger. Fact checking in heterogeneous information networks. In *WWW*, pages 101–102. International World Wide Web Conferences Steering Committee, 2016.

[97] B. Shi and T. Weninger. Scalable models for computing hierarchies in information networks. *Knowledge and Information Systems*, pages 1–31, 2016.

[98] F. Shi, J. G. Foster, and J. A. Evans. Weaving the fabric of science: Dynamic network models of science's unfolding structure. *Social Networks*, 43:73–85, 2015.

[99] M. Sipser. *Introduction to the Theory of Computation*, volume 2. Thomson Course Technology Boston, 2006.

[100] N. A. Smith and M. Johnson. Weighted and probabilistic context-free grammars are equally expressive. *Computational Linguistics*, 33(4):477–491, 2007.

[101] A. Stolcke. An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational linguistics*, 21(2):165–201, 1995.

[102] A. Stolcke and S. Omohundro. Inducing probabilistic grammars by bayesian model merging. In *International Colloquium on Grammatical Inference*, pages 106–118. Springer, 1994.

[103] Z. Sun, H. Wang, H. Wang, B. Shao, and J. Li. Efficient subgraph matching on billion node graphs. *Proceedings of the VLDB Endowment*, 5(9):788–799, 2012.

[104] F. Tao, X. Yu, K. H. Lei, G. Brova, X. Cheng, J. Han, R. Kanade, Y. Sun, C. Wang, L. Wang, and T. Weninger. Research-insight: Providing insight on research by publication network analysis. In *SIGMOD*, pages 1093–1096. ACM, 2013.

[105] R. E. Tarjan and M. Yannakakis. Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs. *SIAM Journal on computing*, 13(3):566–579, 1984.

[106] M. Thoma, H. Cheng, A. Gretton, J. Han, H.-P. Kriegel, A. Smola, L. Song, P. S. Yu, X. Yan, and K. M. Borgwardt. Discriminative frequent subgraph mining with optimality guarantees. *Statistical Analysis and Data Mining*, 3(5):302–318, 2010.

[107] J. Ugander, L. Backstrom, and J. Kleinberg. Subgraph frequencies: Mapping the empirical and extremal geography of large graph collections. In *WWW*, pages 1307–1318. ACM, 2013.

[108] D. J. Watts. Networks, dynamics, and the small-world phenomenon. *American Journal of sociology*, 105(2):493–527, 1999.

[109] D. J. Watts and S. H. Strogatz. Collective dynamics of small-worldnetworks. *nature*, 393(6684):440–442, 1998.

[110] T. Weninger. An exploration of submissions and discussions in social news: mining collective intelligence of reddit. *Social Network Analysis and Mining*, 4(1):1–19, 2014.

[111] T. Weninger, Y. Bisk, and J. Han. Document-topic hierarchies from document graphs. In *CIKM*, pages 635–644. ACM, 2012.

[112] T. Weninger, M. Danilevsky, F. Fumarola, J. Hailpern, J. Han, T. J. Johnston, S. Kallumadi, H. Kim, Z. Li, D. McCloskey, Y. Sun, N. E. TeGrotenhuis, C. Wang, and X. Yu. Winacs: Construction and analysis of web-based computer science information networks. In *SIGMOD*, pages 1255–1258. ACM, 2011.

[113] T. Weninger, F. Fumarola, J. Han, and D. Malerba. Mapping web pages to database records via link paths. In *CIKM*, pages 1637–1640. ACM, 2010.

[114] T. Weninger, F. Fumarola, C. X. Lin, R. Barber, J. Han, and D. Malerba. Growing parallel paths for entity-page discovery. In *WWW*, pages 145–146. ACM, 2011.

[115] T. Weninger, W. H. Hsu, and J. Han. Cetr: content extraction via tag ratios. In *WWW*, pages 971–980. ACM, 2010.

[116] T. Weninger, T. J. Johnston, and M. Glenski. Random voting effects in social-digital spaces: A case study of reddit post submissions. In *Hypertext and Social Media*, pages 293–297. ACM, 2015.

[117] T. Weninger, T. J. Johnston, and J. Han. The parallel path framework for entity discovery on the web. *ACM Transactions on the Web*, 7(3):16, 2013.

[118] T. Weninger, R. Palacios, V. Crescenzi, T. Gottron, and P. Merialdo. Web content extraction: a metaanalysis of its past and thoughts on its future. *ACM SIGKDD Explorations*, 17(2):17–23, 2016.

[119] T. Weninger, X. A. Zhu, and J. Han. An exploration of discussion threads in social news sites: A case study of the reddit community. In *ASONAM*, pages 579–583. IEEE, 2013.

[120] X. Yan and J. Han. gspan: Graph-based substructure pattern mining. In *ICDM*, pages 721–724. IEEE, 2002.

[121] R. Yang, P. Kalnis, and A. K. Tung. Similarity evaluation on tree-structured data. In *SIGMOD*, pages 754–765. ACM, 2005.

[122] M. Yannakakis. Computing the minimum fill-in is np-complete. *SIAM Journal on Algebraic Discrete Methods*, 2(1):77–79, 1981.

[123] Ö. N. Yaveroğlu, T. Milenković, and N. Pržulj. Proper evaluation of alignment-free network comparison methods. *Bioinformatics*, 31(16):2697–2704, 2015.

[124] Z. Yin, M. Gupta, T. Weninger, and J. Han. Linkrec: a unified framework for link recommendation with user attributes and graph structure. In *WWW*, pages 1211–1212. ACM, 2010.

[125] Z. Yin, M. Gupta, T. Weninger, and J. Han. A unified framework for link recommendation using random walks. In *ASONAM*, pages 152–159. IEEE, 2010.

[126] G. U. Yule. A mathematical theory of evolution, based on the conclusions of dr. jc willis, frs. *Philosophical transactions of the Royal Society of London. Series B, containing papers of a biological character*, 213:21–87, 1925.

Tim Weninger

Department of Computer Science and Engineering
353 Fitzpatrick Hall
University of Notre Dame
Notre Dame, IN 46556.

Phone: 574-631-6770
tweninger@nd.edu
http://cse.nd.edu/~tweninge

(a) <u>Professional Preparation</u>

• Kansas State University	Manhattan, KS	Information Systems	B.S., May 2007
• Kansas State University	Manhattan, KS	Computer Science	M.S., Dec 2008
• University of Illinois	Urbana, IL	Computer Science	Ph.D., Aug 2013

(b) <u>Professional Appointments</u>

- September 2013 – present Assistant Professor
  Department of Computer Science and Engineering, University of Notre Dame
- September 2013 – present Affiliated Professor
  Interdisciplinary Center for Network Science and Applications, University of Notre Dame

(c) <u>Products</u>

<u>5 Most closely related</u>

1. Salvador Aguinaga, Rodrigo Palacios, David Chiang and Tim Weninger. Growing Graphs with Hyperedge Replacement Graph Grammars. Proc. of Int. Conf. on Info. and Knowledge Management (**CIKM**'16), Indianapolis, IN 2016.
2. Maria Glenski and Tim Weninger. Rating Effects on Social News Posts and Comments. ACM Trans. Intelligent Systems and Tech. (**TIST**), 2017.
3. Baoxu Shi and Tim Weninger. Fact Checking in Heterogeneous Information Networks. Proc. of 2013 Int. Conf. on the World Wide Web (**WWW**'16), Montreal, Canada. April 2016.
4. Baoxu Shi and Tim Weninger. Scalable Models for Computing Hierarchies in Information Networks. Knowledge and Information Systems (**KAIS**). pp. 1-31, 2016.
5. Salvador Aguinaga, Aditya Nambiar, Zuozhu Liu, Tim Weninger "Concept Hierarchies and Human Navigation"**.** IEEE Conference on BigData (**BigData**'15) Santa Clara, CA. October 29, 2015.

<u>5 Other significant</u>

1. Baoxu Shi and Tim Weninger Discriminative Predicate Path Mining for Fact Checking in Knowledge Graphs. **Knowledge Based Systems**, 104(15), 123-133, 2016.
2. Maria Glenski, Thomas J. Johnston and Tim Weninger "Random Voting Effects in Social-Digital Spaces: A case study of Reddit Post Submissions." ACM Conference on Hypertext and Social Media (**HT**'15), METU, Cyprus, September 1-4, 2015.
3. Baoxu Shi and Tim Weninger. Mining Interesting Meta-Paths from Complex Heterogeneous Information Networks. International Conference on Data Mining (**ICDM**'15) Designing Market of Data, Shenzhen, China, December 14-17, 2014.
4. Tim Weninger, Yonatan Bisk, Jiawei Han. "Document-Topic Hierarchies from Document Graphs." Proc. of Int. Conf. on Info. and Knowledge Management (**CIKM**'12), Maui, Hawaii, Oct. 2012.
5. Tim Weninger, Thomas J. Johnston, Jiawei Han. "The Parallel Path Framework for Entity Discovery on the Web." ACM Transactions on the Web (**TWeb**), 7(3). ACM, 2013.

(d) Synergistic Activities

- Workshop Chair of the 2016 Social Informatics Conference
- Conference Tutorials at WSDM and WWW
- Open Source Software Projects: Dozens of research projects are available at online at https://github.com/nddsg or https://github.com/tweninger
- Developed curricula and course materials for Web Science and Information Retrieval, CSE4/60497 at the University of Notre Dame. Developed course software called the Simple Information Retrieval System (SIRS), which is a complete search engine that focuses on explanation/teaching/readability, available at https://github.com/nddsg/SIRS.
- Journal Reviewer: Transactions on the Web (TWeb), Data Mining and Knowledge Discovery (DAMI), Transactions on Knowledge Discovery and Data Mining (TKDD), Transactions on Intelligent Systems and Technology (TIST), Transactions on Knowledge and Data Engineering (TKDE), Transaction on Information Systems (TOIS), Knowledge and Information Systems (KAIS), Digital Multimedia Broadcasting (IJDMB), Neurocomputing, Information Processing and Management (IPM). Conference Program Committee: KDD, ICDM, IJCAI, SDM, SIGMOD, AAAI, ICML, VLDB, ICDE, WWW, WSDM, CIKM, ASONAM, ADMA, NIPS, IC2S2, NIPS (multiple years).

SUMMARY PROPOSAL BUDGET

YEAR 1

ORGANIZATION **University of Notre Dame**				FOR NSF USE ONLY	
				PROPOSAL NO.	DURATION (months) Proposed / Granted
PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR **Tim Weninger**				AWARD NO.	

A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets)	NSF Funded Person-months CAL	ACAD	SUMR	Funds Requested By proposer	Funds granted by NSF (if different)
1. **Tim Weninger - PI**	0.00	0.00	1.00	11,198	
2.					
3.					
4.					
5.					
6. (0) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE)	0.00	0.00	0.00	0	
7. (1) TOTAL SENIOR PERSONNEL (1 - 6)	0.00	0.00	1.00	11,198	
B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS)					
1. (0) POST DOCTORAL SCHOLARS	0.00	0.00	0.00	0	
2. (0) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.)	0.00	0.00	0.00	0	
3. (2) GRADUATE STUDENTS				42,229	
4. (1) UNDERGRADUATE STUDENTS				1,500	
5. (0) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY)				0	
6. (0) OTHER				0	
TOTAL SALARIES AND WAGES (A + B)				54,927	
C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)				6,874	
TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C)				61,801	
D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING $5,000.)					
TOTAL EQUIPMENT				0	
E. TRAVEL 1. DOMESTIC (INCL. U.S. POSSESSIONS)				5,000	
2. FOREIGN				2,500	
F. PARTICIPANT SUPPORT COSTS					
1. STIPENDS $ 0					
2. TRAVEL 0					
3. SUBSISTENCE 0					
4. OTHER 0					
TOTAL NUMBER OF PARTICIPANTS (0) TOTAL PARTICIPANT COSTS				0	
G. OTHER DIRECT COSTS					
1. MATERIALS AND SUPPLIES				900	
2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION				0	
3. CONSULTANT SERVICES				0	
4. COMPUTER SERVICES				0	
5. SUBAWARDS				0	
6. OTHER				0	
TOTAL OTHER DIRECT COSTS				900	
H. TOTAL DIRECT COSTS (A THROUGH G)				70,201	
I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE) **Modified Total Direct Cost; on campus rate (Rate: 54.5000, Base: 70201)**					
TOTAL INDIRECT COSTS (F&A)				38,260	
J. TOTAL DIRECT AND INDIRECT COSTS (H + I)				108,461	
K. SMALL BUSINESS FEE				0	
L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K)				108,461	

M. COST SHARING PROPOSED LEVEL $ 0 AGREED LEVEL IF DIFFERENT $

PI/PD NAME **Tim Weninger**	FOR NSF USE ONLY		
	INDIRECT COST RATE VERIFICATION		
ORG. REP. NAME*	Date Checked	Date Of Rate Sheet	Initials - ORG

 ***ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET**

SUMMARY PROPOSAL BUDGET

YEAR 2

ORGANIZATION University of Notre Dame				FOR NSF USE ONLY PROPOSAL NO.	DURATION (months) Proposed	Granted
PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR Tim Weninger				AWARD NO.		

A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets)	NSF Funded Person-months CAL	ACAD	SUMR	Funds Requested By proposer	Funds granted by NSF (if different)
1. Tim Weninger - PI	0.00	0.00	1.50	17,302	
2.					
3.					
4.					
5.					
6. (0) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE)	0.00	0.00	0.00	0	
7. (1) TOTAL SENIOR PERSONNEL (1 - 6)	0.00	0.00	1.50	17,302	
B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS)					
1. (0) POST DOCTORAL SCHOLARS	0.00	0.00	0.00	0	
2. (0) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.)	0.00	0.00	0.00	0	
3. (2) GRADUATE STUDENTS				43,496	
4. (1) UNDERGRADUATE STUDENTS				1,500	
5. (0) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY)				0	
6. (0) OTHER				0	
TOTAL SALARIES AND WAGES (A + B)				62,298	
C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)				8,855	
TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C)				71,153	
D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING $5,000.)					
TOTAL EQUIPMENT				0	
E. TRAVEL 1. DOMESTIC (INCL. U.S. POSSESSIONS)				5,000	
2. FOREIGN				2,500	
F. PARTICIPANT SUPPORT COSTS 1. STIPENDS $ 0 2. TRAVEL 0 3. SUBSISTENCE 0 4. OTHER 0					
TOTAL NUMBER OF PARTICIPANTS (0) TOTAL PARTICIPANT COSTS				0	
G. OTHER DIRECT COSTS					
1. MATERIALS AND SUPPLIES				0	
2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION				0	
3. CONSULTANT SERVICES				0	
4. COMPUTER SERVICES				0	
5. SUBAWARDS				0	
6. OTHER				0	
TOTAL OTHER DIRECT COSTS				0	
H. TOTAL DIRECT COSTS (A THROUGH G)				78,653	
I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE) Modified Total Direct Cost; on campus rate (Rate: 54.5000, Base: 78653)					
TOTAL INDIRECT COSTS (F&A)				42,866	
J. TOTAL DIRECT AND INDIRECT COSTS (H + I)				121,519	
K. SMALL BUSINESS FEE				0	
L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K)				121,519	

M. COST SHARING PROPOSED LEVEL $ 0	AGREED LEVEL IF DIFFERENT $		
PI/PD NAME Tim Weninger	FOR NSF USE ONLY INDIRECT COST RATE VERIFICATION		
ORG. REP. NAME*	Date Checked	Date Of Rate Sheet	Initials - ORG

 *ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET

SUMMARY PROPOSAL BUDGET

YEAR 3

ORGANIZATION **University of Notre Dame**				FOR NSF USE ONLY PROPOSAL NO.	DURATION (months) Proposed	Granted
PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR **Tim Weninger**				AWARD NO.		
A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets)	NSF Funded Person-months CAL	ACAD	SUMR	Funds Requested By proposer	Funds granted by NSF (if different)	
1. **Tim Weninger - PI**	0.00	0.00	1.50	**17,821**		
2.						
3.						
4.						
5.						
6. (**0**) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE)	0.00	0.00	0.00	**0**		
7. (**1**) TOTAL SENIOR PERSONNEL (1 - 6)	0.00	0.00	1.50	**17,821**		
B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS)						
1. (**0**) POST DOCTORAL SCHOLARS	0.00	0.00	0.00	**0**		
2. (**0**) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.)	0.00	0.00	0.00	**0**		
3. (**2**) GRADUATE STUDENTS				**44,801**		
4. (**1**) UNDERGRADUATE STUDENTS				**1,500**		
5. (**0**) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY)				**0**		
6. (**0**) OTHER				**0**		
TOTAL SALARIES AND WAGES (A + B)				**64,122**		
C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)				**9,397**		
TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C)				**73,519**		
D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING $5,000.)						
TOTAL EQUIPMENT				**0**		
E. TRAVEL 1. DOMESTIC (INCL. U.S. POSSESSIONS)				**5,000**		
2. FOREIGN				**2,500**		
F. PARTICIPANT SUPPORT COSTS 1. STIPENDS $ **0** 2. TRAVEL **0** 3. SUBSISTENCE **0** 4. OTHER **0**						
TOTAL NUMBER OF PARTICIPANTS (**0**) TOTAL PARTICIPANT COSTS				**0**		
G. OTHER DIRECT COSTS						
1. MATERIALS AND SUPPLIES				**0**		
2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION				**0**		
3. CONSULTANT SERVICES				**0**		
4. COMPUTER SERVICES				**0**		
5. SUBAWARDS				**0**		
6. OTHER				**0**		
TOTAL OTHER DIRECT COSTS				**0**		
H. TOTAL DIRECT COSTS (A THROUGH G)				**81,019**		
I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE) **Modified Total Direct Cost; on campus rate (Rate: 54.5000, Base: 81019)**						
TOTAL INDIRECT COSTS (F&A)				**44,155**		
J. TOTAL DIRECT AND INDIRECT COSTS (H + I)				**125,174**		
K. SMALL BUSINESS FEE				**0**		
L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K)				**125,174**		
M. COST SHARING PROPOSED LEVEL $ **0**	AGREED LEVEL IF DIFFERENT $					
PI/PD NAME **Tim Weninger**				FOR NSF USE ONLY INDIRECT COST RATE VERIFICATION		
ORG. REP. NAME*				Date Checked	Date Of Rate Sheet	Initials - ORG

 *ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET

SUMMARY PROPOSAL BUDGET YEAR 4

ORGANIZATION: **University of Notre Dame**	FOR NSF USE ONLY		
	PROPOSAL NO.	DURATION (months)	
		Proposed	Granted
PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR: **Tim Weninger**	AWARD NO.		

A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets)	NSF Funded Person-months CAL	ACAD	SUMR	Funds Requested By proposer	Funds granted by NSF (if different)
1. Tim Weninger - PI	0.00	0.00	1.00	12,237	
2.					
3.					
4.					
5.					
6. (0) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE)	0.00	0.00	0.00	0	
7. (1) TOTAL SENIOR PERSONNEL (1 - 6)	0.00	0.00	1.00	12,237	
B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS)					
1. (0) POST DOCTORAL SCHOLARS	0.00	0.00	0.00	0	
2. (0) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.)	0.00	0.00	0.00	0	
3. (2) GRADUATE STUDENTS				46,145	
4. (1) UNDERGRADUATE STUDENTS				1,500	
5. (0) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY)				0	
6. (0) OTHER				0	
TOTAL SALARIES AND WAGES (A + B)				59,882	
C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)				8,312	
TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C)				68,194	
D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING $5,000.)					
TOTAL EQUIPMENT				0	
E. TRAVEL 1. DOMESTIC (INCL. U.S. POSSESSIONS)				5,000	
2. FOREIGN				2,500	
F. PARTICIPANT SUPPORT COSTS 1. STIPENDS $ 0 2. TRAVEL 0 3. SUBSISTENCE 0 4. OTHER 0					
TOTAL NUMBER OF PARTICIPANTS (0) TOTAL PARTICIPANT COSTS				0	
G. OTHER DIRECT COSTS					
1. MATERIALS AND SUPPLIES				0	
2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION				0	
3. CONSULTANT SERVICES				0	
4. COMPUTER SERVICES				0	
5. SUBAWARDS				0	
6. OTHER				0	
TOTAL OTHER DIRECT COSTS				0	
H. TOTAL DIRECT COSTS (A THROUGH G)				75,694	
I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE) **Modified Total Direct Cost; on campus rate (Rate: 54.5000, Base: 75694)**					
TOTAL INDIRECT COSTS (F&A)				41,253	
J. TOTAL DIRECT AND INDIRECT COSTS (H + I)				116,947	
K. SMALL BUSINESS FEE				0	
L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K)				116,947	

M. COST SHARING PROPOSED LEVEL $ 0	AGREED LEVEL IF DIFFERENT $

PI/PD NAME: **Tim Weninger**	FOR NSF USE ONLY		
	INDIRECT COST RATE VERIFICATION		
ORG. REP. NAME*	Date Checked	Date Of Rate Sheet	Initials - ORG

 ***ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET**

SUMMARY PROPOSAL BUDGET

YEAR 5

ORGANIZATION **University of Notre Dame**	**FOR NSF USE ONLY**	
	PROPOSAL NO.	DURATION (months) Proposed / Granted
PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR **Tim Weninger**	AWARD NO.	

A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets)	NSF Funded Person-months CAL	ACAD	SUMR	Funds Requested By proposer	Funds granted by NSF (if different)
1. **Tim Weninger - PI**	0.00	0.00	1.00	12,604	
2.					
3.					
4.					
5.					
6. (0) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE)	0.00	0.00	0.00	0	
7. (1) TOTAL SENIOR PERSONNEL (1 - 6)	0.00	0.00	1.00	12,604	
B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS)					
1. (0) POST DOCTORAL SCHOLARS	0.00	0.00	0.00	0	
2. (0) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.)	0.00	0.00	0.00	0	
3. (2) GRADUATE STUDENTS				47,530	
4. (1) UNDERGRADUATE STUDENTS				1,500	
5. (0) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY)				0	
6. (0) OTHER				0	
TOTAL SALARIES AND WAGES (A + B)				61,634	
C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)				8,836	
TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C)				70,470	
D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING $5,000.)					
TOTAL EQUIPMENT				0	
E. TRAVEL 1. DOMESTIC (INCL. U.S. POSSESSIONS)				5,000	
2. FOREIGN				2,500	
F. PARTICIPANT SUPPORT COSTS					
1. STIPENDS $ 0					
2. TRAVEL 0					
3. SUBSISTENCE 0					
4. OTHER 0					
TOTAL NUMBER OF PARTICIPANTS (0) TOTAL PARTICIPANT COSTS				0	
G. OTHER DIRECT COSTS					
1. MATERIALS AND SUPPLIES				0	
2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION				0	
3. CONSULTANT SERVICES				0	
4. COMPUTER SERVICES				0	
5. SUBAWARDS				0	
6. OTHER				0	
TOTAL OTHER DIRECT COSTS				0	
H. TOTAL DIRECT COSTS (A THROUGH G)				77,970	
I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE) **Modified Total Direct Cost; on campus rate (Rate: 54.5000, Base: 77970)**					
TOTAL INDIRECT COSTS (F&A)				42,494	
J. TOTAL DIRECT AND INDIRECT COSTS (H + I)				120,464	
K. SMALL BUSINESS FEE				0	
L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K)				120,464	

M. COST SHARING PROPOSED LEVEL $ 0 AGREED LEVEL IF DIFFERENT $

PI/PD NAME **Tim Weninger**	**FOR NSF USE ONLY**		
	INDIRECT COST RATE VERIFICATION		
ORG. REP. NAME*	Date Checked	Date Of Rate Sheet	Initials - ORG

 ***ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET**

SUMMARY PROPOSAL BUDGET

Cumulative

ORGANIZATION: University of Notre Dame				FOR NSF USE ONLY	
				PROPOSAL NO.	DURATION (months): Proposed / Granted
PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR: Tim Weninger				AWARD NO.	

A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets)	NSF Funded Person-months CAL	ACAD	SUMR	Funds Requested By proposer	Funds granted by NSF (if different)
1. Tim Weninger - PI	0.00	0.00	6.00	71,162	
2.					
3.					
4.					
5.					
6. ( ) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE)	0.00	0.00	0.00	0	
7. (1) TOTAL SENIOR PERSONNEL (1 - 6)	0.00	0.00	6.00	71,162	
B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS)					
1. (0) POST DOCTORAL SCHOLARS	0.00	0.00	0.00	0	
2. (0) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.)	0.00	0.00	0.00	0	
3. (10) GRADUATE STUDENTS				224,201	
4. (5) UNDERGRADUATE STUDENTS				7,500	
5. (0) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY)				0	
6. (0) OTHER				0	
TOTAL SALARIES AND WAGES (A + B)				302,863	
C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)				42,274	
TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C)				345,137	
D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING $5,000.)					
TOTAL EQUIPMENT				0	
E. TRAVEL 1. DOMESTIC (INCL. U.S. POSSESSIONS)				25,000	
2. FOREIGN				12,500	
F. PARTICIPANT SUPPORT COSTS					
1. STIPENDS $ 0					
2. TRAVEL 0					
3. SUBSISTENCE 0					
4. OTHER 0					
TOTAL NUMBER OF PARTICIPANTS (0) TOTAL PARTICIPANT COSTS				0	
G. OTHER DIRECT COSTS					
1. MATERIALS AND SUPPLIES				900	
2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION				0	
3. CONSULTANT SERVICES				0	
4. COMPUTER SERVICES				0	
5. SUBAWARDS				0	
6. OTHER				0	
TOTAL OTHER DIRECT COSTS				900	
H. TOTAL DIRECT COSTS (A THROUGH G)				383,537	
I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE)					
TOTAL INDIRECT COSTS (F&A)				209,028	
J. TOTAL DIRECT AND INDIRECT COSTS (H + I)				592,565	
K. SMALL BUSINESS FEE				0	
L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K)				592,565	
M. COST SHARING PROPOSED LEVEL $ 0	AGREED LEVEL IF DIFFERENT $				

PI/PD NAME: Tim Weninger	FOR NSF USE ONLY		
	INDIRECT COST RATE VERIFICATION		
ORG. REP. NAME*	Date Checked	Date Of Rate Sheet	Initials - ORG

C *ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET

Budget Justification

Salaries and Wages

Salaries and wages (including stipends) are based on the University of Notre Dame's budget forecast for fiscal year 2016-2017 with 3 percent (3%) annual increment compensation.

Senior Personnel: Six months summer salary is requested for PI Weninger; summer salary is allocated according to existing grants and expectation for future funding.

Other Personnel: Support ($27,333 first year stipend) for two Computer Science Graduate Students (9 months/year) is requested for all five years. Personnel budget lines include a three per cent salary escalation to accommodate a portion of anticipated merit raises during the project period.

Support ($1,500 per year) for a Computer Science undergraduate student is requested for all five years for the education and outreach program. Although not included in this budget, the PI will apply to the REU program to further supplement the education and outreach objectives of this proposal.

Personnel responsibilities

Senior Personnel: The PI (Weninger) will direct the two graduate students and will apply for REU funding for undergraduate students.

Non-Senior Personnel: Two graduate students from computer science will provide the design, algorithm development, and software engineering effort to implement the proposed work.

Fringe Benefits [FY17]

Employee benefits are directly charged as a percentage of salaries and wages. The regular faculty rate is 26.1% with a inflation factor of 1.01%, the graduate student rate is 9.3% with a inflation factor of 1.05%, the undergraduate student rate is 1.6% with no inflation factor. Fringe rate percentages are negotiated annually. These rates are applied throughout the life of the project.

Travel (domestic and foreign)

Funds are requested to cover travel costs necessary for the PI and his graduate students to meet with project collaborators, to attend workshops organized by the team as part of the educational outreach plan, and to participate in conferences and present the results of this research. These conferences include: ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD), World Wide Web Conference (WWW), IEEE Conference on Data Mining (ICDM), ACM Conference on Information and Knowledge Management (CIKM), etc. The requested funds will support the PI and his graduate students to attend four domestic conferences per year and one international conference each year. For domestic travel, the estimated average cost per trip is $1,250 for four travel days, which includes airfare, ground transportation, lodging, meals, and conference registration. For foreign travel, the estimated cost is $2,500 per six travel days.

Computer Services

$0 is requested for computer services. The PI has four large servers purchased with startup funds: each server has 256GB of RAM memory, 6 TB of available disk, and 64 processing cores. Further detailed can be found in the Facilities, Equipment and Other Resources document.

Materials and Supplies

A modest amount ($900) is requested for Materials and Supplies to cover the cost of expendable, project related materials such as photocopying and the printing of large posters.

Indirect Costs

Notre Dame's federally-negotiated indirect cost rate is applied to modified total direct costs (all direct costs except participant support, equipment, and the first $25,000 of each sub-award). (U.S. Department of Health and Human Services is Notre Dame's cognizant federal agency; the F&A rate agreement is dated October 14, 2015, and is applicable for five years.) For this project, the F&A rate of 54.5% for research conducted on campus is used throughout the life of the project.

Current and Pending Support
Weninger, Timothy
University of Notre Dame

Current

Title:	Socio-Digital Influence Attack Models and Deterrance (FA9550-15-1-0003.2)
PI:	Weninger, Timothy
Source:	Department of the Air Force
Total Award Amount:	$349,379
Grant Period:	12/01/2014 - 11/30/2017
Location of Project:	University of Notre Dame
Person Months	2.0

Pending

Title:	CAREER: Principled Structure Discover for Network Analysis (this proposal)
Source:	NSF
Amount:	$592,565
Grant Period:	05/01/2017 - 04/30/2022
Location of Project:	University of Notre Dame
Person Months	1.0

Title:	Mining Conversation Trails for Effective Group Behavior
Source:	Department of Army
Amount:	$359,988
Grant Period:	01/01/2017 - 12/31/2019
Location of Project:	University of Notre Dame
Person Months	0.5

Facilities, Equipment and Other Resources

The PI has access to the following facilities, equipment, and other resources, which will be adequate to pursue the proposed research goals. Wherever appropriate, this document will describe how the existing resources will be used for the proposed research project. Specifically, the PI involvement in other ongoing large-scale studies, university centers and data collection efforts has equipped him with the laboratory space, tools, and equipment necessary for this project.

Resources of Weninger's Lab

Researchers on this project reside in the the large and modern Interdisciplinary Center for Network Science and Applications (iCeNSA) Lab, which is a short walk from the PI's office. The PI is also located within a short walk through the campus of Notre Dame from the Center of Research Computing, which will house and maintain the collected data sets. The laboratory space will be used for the data collection programming and analysis of the collected data.

Personal Computers The University of Notre Dame provides graduate students with computers and monitors at their request upon their arrival. The university also provides a new personal laptop or desktop computer to faculty once every three years. Ancillary devices such as printers, keyboards, etc. are provided by the CSE department at no charge.

Computing Servers The Notre Dame Data Science Group (ND-DSG), under the direction of the PI, currently owns adequate computing equipment required for this project. This includes four Dell PowerEdge (R815) servers with 256GB of RAM memory, 6 TB of available disk, and 64 processing cores each.

Because of this existing infrastructure the PI does not require equipment funding.

Software The ND-DSG has access to all of the software required for this project, including Matlab, Python, Java programming environments as well as Hadoop, HDFS, and Spark systems configured on the computing servers.

College-wide Computing Resources

All researchers in the College of Engineering at the University of Notre Dame have easy access to the Engineering College's Workstation Cluster, consisting of more than 100 state-of-the-art workstations. The workstations and computers are all connected to the Internet network, allowing access to a nationwide repository of data and programs. Student developers participating in this project will have anytime access to these resources.

Center for Research Computing (CRC)

The CRC houses two complementary resources: the High Performance Computing section, providing over 8000 cores of computational power with the associated support infrastructure for both hardware and installed software, and the Cyberinfrastructure section, empowering faculty and industry partners to develop research environments that support advanced data and information processing services including acquisition, storage, management, integration, mining and visualization of data.

These services include: the utilization of up to 10TB of redundant distributed (network) storage, nightly off-site backup, SVN source code repositories, and basic web services to share data sets residing in said storage. Security of the data is ensured through centralized institutional authentication controls and encrypted data storage.

All system administration is carried out by CRC with access provided only to students who are members of the PI's research group. All subject logins and data inputs (via the web or mobile) are strictly protected by digital certificates (public/private keys) ensuring appropriate confidentiality for all communications to the server. Database and storage array access are strictly separated from the general Internet by virtue of tightly regulated firewall rules.

For more information about CRC policies and services please see: `http://crc.nd.edu/index.php/aboutcrc/policies`.

Data Management Plan

Data & Software

The types of data that will be consumed and produced in this research include a wide range of graph datasets, examples of which are given in the Project Description. The datasets will primarily be obtained through the Internet from publicly available sources such as private institutions, government agencies and academic institutions and stored locally at facilities operated by the University of Notre Dame Center for Research Computing (CRC). There is a need for significant storage space requirements in this project. For example, a single Wikipedia snapshot database is approximately 50GB. Many such datasets, and potentially daily versions thereof, as well as an array of similar types of data from other sources will be used throughout this project. In total, approximately 10TB of storage will be needed over the course of the project, which will be physically stored on existing servers, hosted by Notre Dame's CRC, and managed by the PI.

Several software packages will also be developed to process, and analyze these datasets. The majority of the processing will be performed using basic shell, Perl, Python or similar scripting languages. The analysis will rely on mathematical and statistical software packages such as Matlab and R, as well as custom codes written in C/C++, Python, Java, etc. Visualizations will be produced using Matlab's plotting capabilities. Many of these software packages are freely available; and Notre Dame provides access to all of these and required commercial software via a campus license.

Data Standards

Datasets will come in heterogeneous formats ranging from ASCII (e.g., plain-text, comma-separated values) to complex XML data and formats specifically designed for graph data (*e.g.*, nodeXL, dot, rdf). All of these are open formats and built-in routines and/or external libraries exist in many popular software packages to access them; no proprietary software will be required. One convenient feature of the widely used data formats is the ability to include metadata, which is extensively used in the scientific computing and especially network science communities. In addition, the binary formats are highly compressed relative to their plain-text representation, so such data formats will be used for long-term storage.

Access and Sharing

The PI is committed to open science, data sharing and open source software licences.

Data The PI will establish a public repository of network datasets. The access policy will be that at the time of publication, datasets used or produced will be made available under the CC-BY license free of charge. Files which may exceed the capacity of the Web server will be provided to interested parties upon request. The PI is committed to providing convenient data access to a broad user base. Internally, data may be kept on hard drives for short-term storage and analysis.

Source Code Source code will be made publicly available on GitHub under the CC-BY license, and the PI will accept changes and updates offered by the community, *i.e.*, pull requests. Software codes and intermediate datasets at Notre Dame will be protected from unauthorized access by firewall and password authentication.

Publications Except under special circumstances, the PI will submit to publication venues that allow open access or submission to ArXiV. Therefore, scholarly publications will be posted to ArXiV upon acceptance under the CC-BY license.

Re-Use, Re-Distribution, and Derivatives

Data will be provided "as is" via general access including metadata information in README text files or via the metadata capabilities of scientific file formats. Under the terms of the CC-BY license, use of the data in other publications or products will be permitted if the source is acknowledged. Data and data products will not be copyrighted.

Archiving

The analysis will be performed at Notre Dame, where data is stored on hard drives in RAID configuration and automatically archived to tertiary storage to ensure preservation. Final data sets will be archived at Notre Dame, which has an automatic procedure for permanent archival of massive data sets. The PI's goal is to permanently maintain the public source code and data repository pending availability of funds beyond the duration of this project.

UNIVERSITY OF
NOTRE DAME

Department of Computer Science and Engineering

384 Fitzpatrick Hall
Notre Dame, Indiana
46556 USA

Kevin W. Bowyer
Schubmehl-Prein Professor and Department Chair

tel **(574) 631-9978**
fax **(574) 631-9260**

June 22, 2016

To the National Science Foundation Review Panel,

I am writing to express my enthusiastic support for Professor Tim Weninger's proposal entitled "**CAREER: Principled Structure Discovery for Network Analysis**" to the National Science Foundation. Professor Weninger is a prominent researcher, one of our star young faculty members in Computer Science & Engineering, who plays an important role in our plans to expand our research and teaching activities in Data Science and as part of the Notre Dame's Interdisciplinary Center for Network Science and Applications.

Having hired Prof. Weninger after a national search, Notre Dame is investing in Tim's career development. In particular we have:

1) Provided a startup package, which includes support for two full time graduate students over three full years as well as some summer salary.
2) Provided office space and lab space for Prof. Weninger's graduate students.
3) Purchased computing equipment that is currently housed at Notre Dame's state-of-the-art Center for Research Computing facility.
4) Covered travel expenses for Prof. Weninger to attend, and in some cases, deliver invited talks at the US National Academies, NSF, ASEE's National Excellence in Teaching Institute, and the Computing Research Association.
5) Assigned Professor Nitesh Chawla, as a mentor to Tim. Professor Chawla, a PI of several NSF, ARO, AFOSR, and ONR awards, will provide guidance to Professor Weninger on research, education, and outreach activities.

In addition the above commitments, the university, college and department have many fellowships and graduate assistantships available to graduate students, thereby supplementing their support. Furthermore, neither the university nor the graduate school charge tuition to sponsored graduate students. As a result, the university covers much of the cost associated the graduate student workers associated with this proposal.

Prof. Weninger teaches one course per semester, and receives course evaluations that are in the top 25% of the department. Depending on the timing of this award, Prof Weninger will be considered for promotion and tenure after year three of this project.

Prof Weninger has been a frequent sponsor in NSFs research experience for undergraduates (REU), and has sponsored two high school teachers via NSFs research experience for teachers (RET) program in the summer of 2016.

I verify that Professor Weninger is eligible for the NSF CAREER award. Professor Weninger holds a full-time, tenure-track faculty appointment in the Department of Computer Science and Engineering track at the University of Notre Dame. Professor Weninger received his doctoral degree in Computer Science from the University of Illinois Urbana-Champaign in 2013. Furthermore, Prof. Weninger is a US Citizen and is eligible for consideration of the PECASE award.

I urge the most serious consideration of Professor Weninger's proposal. If I can provide any additional information, please do not hesitate to contact me.

Sincerely,

Kevin W Bowyer

Kevin W. Bowyer.
Schubmehl-Prein Professor
Chair, Department of Computer Science and Engineering
University of Notre Dame

John Parkhill
251 Nieuwland Science Hall
Notre Dame, Indiana, 46556

UNIVERSITY OF NOTRE DAME

Phone: (617) 642-3837
e-mail: john.parkhill@gmail.com

June 10th, 2016

Dear Members of the Review Panel:

If the proposal submitted by Tim Weninger entitled "CAREER: Principled Structure Discovery for Network Analysis" is selected for funding by the NSF, it is my intent to collaborate and/or commit resources as detailed in the Project Description.

Sincerely,

John Parkhill

THE UNIVERSITY OF CHICAGO
DEPARTMENT OF SOCIOLOGY
1126 EAST 59TH STREET
CHICAGO • ILLINOIS 60637
http://sociology.uchicago.edu

June 10, 2016

National Science Foundation
United States of America

Dear Review Panel and ad hoc reviewers:

If the proposal submitted by Dr. Tim Weninger entitled *CAREER: Principled Structure Discovery for Network Analysis* is selected for funding by the NSF, it is my intent to collaborate and/or commit resources as detailed in the Project Description.

Sincerely,

James A. Evans
Professor of Sociology
University of Chicago
jevans@uchicago.edu
(773)834-3612
http://home.uchicago.edu/~jevans

UNIVERSITY OF
NOTRE DAME
College of Engineering
Computer Science & Engineering

384 Fitzpatrick Hall Notre Dame, Indiana 46556-5637 USA
tel (574) 631-8320 *fax* (574) 631-9260 *email* cse@cse.nd.edu

National Science Foundation
Arlington, VA

2016/7/18

Dear Members of the Review Panel:

If the proposal submitted by Tim Weninger entitled "CAREER: Principled Structure Discovery for Network Analysis" is selected for funding by the NSF, it is my intent to collaborate and/or commit resources as detailed in the Project Description.

Sincerely,

David Chiang

David Chiang
Associate Professor
Department of Computer Science and Engineering
University of Notre Dame

NIRSEF
Northern Indiana Regional Science and Engineering Fair

Dear Members of the Review Panel:

If the proposal submitted by Tim Weninger entitled "CAREER: Principled Structure Discovery for Network Analysis" is selected for funding by the NSF, it is my intent to collaborate and/or commit resources as detailed in the Project Description.

Sincerely,

Alisa Zornig Gura

Alisa Zornig Gura
NIRSEF Executive Director

SOUTH BEND COMMUNITY SCHOOL CORPORATION

215 South St. Joseph Street South Bend, Indiana 46601

Telephone (574) 283-8000

National Science Foundation
Arlington, VA

Dear Members of the Review Panel:

If the proposal submitted by Tim Weninger entitled “CAREER: Principled Structure Discovery for Network Analysis” is selected for funding by the NSF, it is my intent to collaborate and/or commit resources as detailed in the Project Description.

Sincerely,

George Logsdon
Math Chair
Riley High School
South Bend, IN

UNIVERSITY OF
NOTRE DAME

KANEB CENTER FOR TEACHING AND LEARNING

353 DeBartolo Hall
Notre Dame, Indiana
46556-5602

telephone **(574) 631-9146**
fax **(574) 631-8047**
email: **kaneb.2@ND.EDU**
Website: Kaneb.nd.edu

July 18, 2016

National Science Foundation
4201 Wilson Boulevard
Arlington, VA 22230

Dear Members of the Review Panel:

If the proposal submitted by Dr. Timothy Weninger entitled "**CAREER: Principled Structure Discovery for Network Analysis**" is selected for funding by the NSF, it is my intent to collaborate and/or commit resources as detailed in the Project Description.

Sincerely,

Daniel J. Hubert, PhD
Associate Director for Learning Outcomes Assessment