AFRL-OSR-VA-TR-2014-0070

(DARPA) Algebraic Statistics for Network Models

SONJA PETROVIC

PENNSYLVANIA STATE UNIVERSITY

02/19/2014
Final Report

DISTRIBUTION A: Distribution approved for public release.

AIR FORCE RESEARCH LABORATORY
AF OFFICE OF SCIENTIFIC RESEARCH (AFOSR)/RSL
ARLINGTON, VIRGINIA 22203
AIR FORCE MATERIEL COMMAND

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE *(DD-MM-YYYY)* 02-14-2014	**2. REPORT TYPE** final	**3. DATES COVERED** *(From - To)* 7/15/2012-1/14/2014

4. TITLE AND SUBTITLE Algebraic Statistics for Network Models	**5a. CONTRACT NUMBER**
	5b. GRANT NUMBER FA9550-12-1-0392
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Petrovic, Sonja Rinaldo, Alessandro Fienberg, Stephen E.	**5d. PROJECT NUMBER**
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) The Pennsylvania State University	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) USAF, AFRL AF OFFICE OF SCIENTIFIC RESEARCH 875 NORTH RANDOLPH STREET, RM 3112 ARLINGTON, VA 22203	**10. SPONSOR/MONITOR'S ACRONYM(S)** AFOSR & DARPA
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

UNL

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This project focused on the family of exponential random graph models (ERGMs) for networks, characterized by global network summary statistics. These models are especially attractive because they are built precisely around the kinds of network characteristics that analysts are concerned with in most practical applications. The team has proposed a systematic program of mathematical research into the algebraic geometric structure of parameter estimation and assessing model fit for these and related statistical models.

The team has reached all three of the proposed Phase I measurable milestones and made significant progress toward future proposed work, reaching an additional milestone originally proposed for later phases.

In particular, the team has: (1) created new tools for assessing the goodness of fit of models and comparison of models within the ERGM class; and (2) characterized the statistical properties of ERGMs using geometric tools and, specifically, identified when ERGMs are "nice," i.e., not exhibiting near-degeneracies of the sort described in the statistical literature.

15. SUBJECT TERMS

Model/data fit, goodness of fit, random graph, social network, algebraic statistics, MLE existence, hypergraph, toric ideal, comput

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** John W. Hanold
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	UU	1	**19b. TELEPHONE NUMBER** *(include area code)* 814-865-1372

Final report for DARPA GRAPHS Phase I
Algebraic Statistics for Network Models
FA9550-12-1-0392

Sonja Petrović
petrovic@psu.edu[1]
Department of Statistics
Pennsylvania State University

Alessandro Rinaldo
arinaldo@cmu.edu
Department of Statistics
Carnegie Mellon University

Stephen E. Fienberg
fienberg@stat.cmu.edu
Department of Statistics, Heinz College, Machine Learning Department, Cylab
Carnegie Mellon University

1. Abstract

This project focused on the family of exponential random graph models (ERGMs) for networks, characterized by global network summary statistics. These models are especially attractive because they are built precisely around the kinds of network characteristics that analysts are concerned with in most practical applications. The team has proposed a systematic program of mathematical research into the algebraic geometric structure of parameter estimation and assessing model fit for these and related statistical models. The team has reached all three of the proposed Phase I measurable milestones and made significant progress toward future proposed work, reaching a major part of an additional milestone originally proposed for later phases. In particular, the team has: (1) created new tools for assessing the goodness of fit of models and comparison of models within the ERGM class; and (2) characterized the statistical properties of ERGMs using geometric tools and, specifically, identified when ERGMs are nice, i.e., not exhibiting near-degeneracies of the sort described in the statistical literature. The project resulted in 17 publications and preprints, and included two postdoctoral researchers and a graduate student (for one semester).

[1] Current address: Department of Applied Mathematics, Illinois Institute of Technology, Chicago IL 60616. Current email: Sonja.Petrovic@iit.edu

2. Personnel

The Team consists of Sonja Petrović (PI), Alessandro Rinaldo and Stephen E. Feinberg. The Team has hired two postdoctoral researchers on the grant: Despina Stasi and Kayvan Sadeghi. In addition, during the last five months of the project, there was one graduate student working under Petrović, supported by the grant for one semester.

3. Executive Summary of Progress

This project addressed several major challenges that exist as obstacles for accurate modeling and inference for network structures. While most of the literature addresses fast- yet heuristic- approaches for summarizing network characteristics, our Team's focus has been on providing the formalism and theoretical framework necessary for providing increased confidence in the applied work, and tools for proving its relevance and validity. In particular, we provide statistical tools for fitting network models, for modeling edge dependencies in networks that naturally occur in many types of network data, and develop a family of new model hierarchies for higher-order interactions.

We have addressed majority of the tasks and research aims that were specified in Phase I of the original research proposal, and have also addressed some research aims proposed for later phases, since the work evolved in that direction naturally. As per the executive summary in the project proposal, we have made significant steps toward the following goals:

- Characterizing the statistical properties of ERGMs using geometric tools and, specifically, identify when ERGMs are "nice," i.e., not exhibiting near-degeneracies of the sort described in the statistical literature.
- Develop methods for generating conditional distributions for selected models given their minimal sufficient statistics for large number of nodes.
- Create new tools for assessing the goodness of fit of models and comparison of models within the ERGM class and, in particular, extend our algebraic statistics and Markov bases results from p_1 to the broader class of ERGMs.
- Continue to develop new algebraic tools for hypergraphs and use them for extensions to ERGM models when nodes are grouped, as well as linked via directed or undirected ties.

These four aspects of our work encompass four of the six major directions our the proposed research for this project, in all three phases.

4. Specific knowledge advances resulting from this project

The problems we address are extremely general, as we are developing a method, principle to do inference, which leads to a variety of different problems. Specifically, as a result of our project's Phase I, we now have a good understanding of the following fundamental statistical problems for network data.

1. Existence and scaling of maximum likelihood estimators (Petrović-Rinaldo-Fienberg)
2. Comparison between MLE and pseudo-MLE (Rinaldo-Fienberg)

3. Consistency under sampling (Rinaldo)
4. Goodness of fit for problem and sampling from conditional distributions (Stasi-Petrović joint with entire team)
5. Network data privacy (Fienberg)
6. Spectral clustering (Rinaldo)
7. Modeling (in)dependence structures in networks (Sadeghi)
8. Beta hypergraph models (Stasi joint with entire team)
9. Modeling framework for dk-graphs (Sadeghi joint with entire team).

As a result of this effort, we can thus answer several questions that were previously unanswerable in a systematic manner. Below we highlight some of the key questions we are now able to answer using algebraic statistics and geometry.

Question 1: Can the MLE be considered as a satisfactory generative model for the data at hand?

Answer: This is a basic question tied to a broader problem of determining if a particular log-linear model fits an observed social network. We use algebra, combinatorics and Markov bases to give a constructing way of answering this question for ERGMs of interest.

Question 2: How do we model edge-dependencies in a social network / random graph?

Answer: This question is tied to a broader problem: Is the assumption of independent edges (dyads) a good one? We extend the basic modeling framework and propose a new model hierarchy, for which we can show it has good statistical properties.

Question 3: How do we model higher-order relations in a statistically consistent way?

Answer: From a broader perspective, we are questioning the usual, often assumed reasonable, data/modeling requirement: distinguishing between binary and k-ary relations between groups of people. In particular, we consider if there is information lost when we consider "just" the underlying graph. The answer is "yes". We propose a new family of hypergraph-based models that extend the standard random graph models, and provide the relevant statistical theory for analyzing them.

5. Milestones

There were three milestones in Phase I of the proposed project. These have been reached, as specified below.

Phase 1, Month 9 milestone: Determination of the model polytopes for general network models and their properties (facial sets and normal fan, when appropriate). This research goal will be pursued quite broadly for general network models, even though special emphasis will be put on the following models: p_1 models and their variants, Markov graph models and generalized Markov graph models and hypergraphs.
(One sub-task which remains to be completed: explicit, implementable determination of the polytopes for Markov graph models. this particular sub-task is currently work in progress.)

Phase 1, Month 14 milestone: Determination of the commutative algebra properties of toric ideals associated to hypergraphs, focusing mostly on minimal generating sets and Gröbner basis for these toric ideals, which will form the basis for all relevant statistical computations. *A sub-milestone:* a generative description of the moves for hierarchical families of hypergraphs.

Phase 1, Month 18 milestone: Characterization of the pseudo-likelihood and comparison of its existence to that of the MLE. Detailed comparison of the geometry of pseudo-MLE with that of MLE.

Additionally, significant progress and proof-of-concept algorithms have been implemented toward the major project-wide milestone that is supposed to occur in the last Phase of the entire four-year effort. Namely, we have done an initial implementation of the following, for the p_1 random graph model for directed graphs with reciprocation:

Phase 3, Month 44 milestone: Implementation of the goodness-of-fit test for network models, including an exhaustive data-analytic investigation of the data set.
Plan going forward: As this is one of the major milestones proposed in our four-year effort, with continued support the Team plans to continue to work on extending this for larger datasets. In particular, we will seek collaboration with other GRAPHS teams, specifically Notre Dame and South Carolina, to help us reach the goal in terms of implementation. More theoretical work also remains to be done to ensure that any implementations will be statistically significant, relevant, and provably correct. The work we have done during Phase I has provided the needed foundational groundwork that will help us complete this entire effort.

6. Manuscripts

What follows is a description of advances that have taken place since 15 July 2012. In terms of dissemination, we only list activities undertaken by the Team, and not other collaborators on the products listed.

The team has completed 12 manuscripts, some of which are already published or in print, and have at least 4 more that are well underway.

1. Rinaldo, A., Petrović, S. and Fienberg, S.E. Maximum Likelihood Estimation in the Beta Model. *Annals of Statistics* 2013, Vol. 41, No. 3, 1085-1110.

   - Related proposed research direction: **Research Aim 1**.

2. Rinaldo, A., Petrović, S. and Fienberg, S.E. How Does Maximum Likelihood Estimation for the p_1 Model Scale for Large Sparse Networks?. Submitted paper for *poster presentation at NIPS 2012.*

   - Related proposed research direction: **Research Aim 1**.

3. Yang, X., Rinaldo, A. and Fienberg, S.E. (2013). Dependent Exponential Random Graph Models and Comparison between Maximum Likelihood Estimation and Maximum Pseudo Likelihood Estimation, submitted to the *Journal of Algebraic Statistics.*

   - Related proposed research direction: **Research Aim 2**.

4. Shalizi, C. R. and Rinaldo, A. (2012). Consistency under Sampling of Exponential Random Graph Models. *Annals of Statistics*, to appear.

    - This work fits under the **broader goal** of the grant, which is characterizing statistical properties of ERGMs. Although no specific task specified the development of this manuscript, it was inspired by interactions with other GRAPHS teams and realization that the theoretical underpinnings of consistency were inadequate, while such properties were often assumed in the literature. The work address this issue and explains when the property does not hold.

5. Garcia-Puente, L., Petrović, S. and Sullivant, S. GraphicalModels in Macaulay2. *Journal of Software for Algebra and Geometry*, (2013) Volume 5.

    - Related proposed research direction: **Research Aim 8**. This work does not study network models directly, but the tools developed here will be used (directly) on network models. This software package applies more generally.

6. Fienberg, S.E. A Brief History of Statistical Models for Network Analysis and Open Challenges. *Journal of Computational and Graphical Statistics* **21** (4), (2012) 825–839. DOI:10.1080/10618600.2012.738106.

    - This work fits under the **broader goal** of characterizing statistical properties of network models.

7. Sadinle, M. The Strength of Arcs and Edges in Interaction Networks: Elements of a ModelBased Approach. Technical report, 1/21/2013, Carnegie Mellon University. [The author is a PhD student working with S.E.Fienberg at Carnegie Mellon.]

8. Gross, E. and Petrović, S. A combinatorial degree bound for toric ideals of hypergraphs. *International Journal of Algebra and Computation*, Volume 23, Issue 06, pp. 1503-1520, September 2013.

    - Related proposed research direction: **Research Aim 6**.

9. Fienberg, Stephen E. (2013) "Is the Privacy of Network Data an Oxymoron?," Journal of Privacy and Confidentiality: Vol. 4: Iss. 2, Article 1.
   Available at: http://repository.cmu.edu/jpc/vol4/iss2/1

    - Related proposed research direction: the **general goal** to provide a systematic understanding of properties of network models.

10. Petrović, S. and Stasi, D. Toric algebra of hypergraphs. *Journal of Algebraic Combinatorics*, February 2014, Volume 39, Issue 1, pp 187-208. (DOI 10.1007/s10801-013-0444-y)

    - Related proposed research direction: **Research Aim 5**.

11. Lei, J. and Rinaldo, A. (2013). Consistency of Spectral Clustering in Sparse Stochastic Block Models. Preprint (http://arxiv.org/abs/1312.2050)

    - This one is among the few additional research activities we list in this report that are directly relevant to the **broader goals** of the research performed under the grant. The work analyzes the performance of a practical spectral clustering algorithm for community extraction in the stochastic block model, which is of interest for the proposed research aims.

12. Sadeghi, K. (2013). Stable mixed graphs. Bernoulli 19 2330-2358.
    And:
13. Sadeghi, K., and Lauritzen, S. L. Markov properties for mixed graphs. *Bernoulli*, to appear.
    - Related proposed research direction: **Research Aim 8**. This work does not study network models directly, but the tools developed here will be used (directly) on network models and are directly relevant for our research goals.
14. Sadeghi, K. and Rinaldo, A. Statistical models for degree and bi-degree distributions of networks. Submitted to ACML

15. Sadeghi, K., Marginalization and conditioning for LWF chain graphs, Submitted to the Annals of Statistics

16. Yin, M., Rinaldo, A. and Fadnavis S. (2013). Asymptotic quantization of exponential random graphs. Preprint (http://arxiv.org/abs/1311.1738)

17. Gross, E., Petrović, S. and Stasi, D. Dynamic algorithm for generating Markov moves for the p_1 model. Preprint.
    - Related proposed research direction: **Research Aims 4, 5, 6**.

7. Presentations, further dissemination, and synergistic activities

The Team has been involved in numerous presentations and synergistic activities during the period of the grant. Below, we only list those activities led by the team members, and not other collaborators involved in the products. The chronological list below includes the following information: presenter; the topic of presentation/meeting; occasion/location of presentation. References to manuscript numbers above are listed in place of topics, where appropriate.

06/29/12 Rinaldo. Manuscript 1. Invited lecture at the Joint Statistical Meeting in San Diego, in an invited session on network models.

10/01/12 Petrović. Manuscript 1. Undergraduate Colloquium, Loyola University Chicago.

12/04/12 Petrović. Manuscripts 8, 10. Combinatorial Commutative Algebra and Applications Workshop, MSRI, Berkeley CA.

12/08/12 Rinaldo. Manuscript 1. Poster presentation at NIPS 2012 session on networks.

12/11/12 Rinaldo. Lecture on “Maximum Likelihood Estimation in Exponential Families for Discrete Data”. Stanford Statistics Department Colloquium

02/28/13 Petrović. Manuscripts 1, 10. Invited lecture at the Minisymposium on “Analysis and Modeling of Static and Dynamic Networks” at the SIAM Conference on Computational Science & Engineering. Boston, MA.

03/16/13 Stasi. Manuscript 10. Poster presentation at the 2013 AWM Research Symposium. Santa Clara, CA.

03/20/13 Petrović. Manuscripts 8, 10. Invited lecture at the Workshop on 'Perspectives and Emerging topics in Algebra and Kombinatorics'. Haus Bergkranz, Riezlern (Kleinwalsertal), Austria.

Software Petrović. Manuscript 5 - related software release. As of April 2013, the fully documented package `GraphicalModels.m2` is included in the current release of the software Macaulay2. The package includes small graphical model examples for every function.

06/06/13 Petrović. Manuscripts 8, 10. Invited lecture at the Scientific Session on Commutative Algebra and Combinatorics at the Canadian Mathematical Society Summer Meeting. Dalhousie University and Saint Mary's University, Halifax, Nova Scotia, Canada.

06/19/13 Sadeghi. "Hierarchical graphical independence models for networks", presented at the Exponential random network models workshop at the American Institute of Mathematics in Palo Alto.

08/03/13 Fienberg. "Some Open Challenges in the Analysis of Network Data," Joint Statistics Meetings, Montreal Canada.

09/16/13 Fienberg. "Challenges of Mobile Trajectory Data: Location and time-Aware Data Analytics," 48th Meeting of Asian-Pacific Economic Cooperation Telecommunications and Information Working Group, Honolulu, HI.

09/23/13 Sadeghi. Manuscript 13. IIT Applied Mathematics Colloquium, Chicago.

10/05/13 Fienberg. "New Multivariate Challenges and Privacy Issues in the Context of Social/Commercial Systems Research," P.R. Krishnaiah Memorial Lecture, Department of Statistics, Pennsylvania State University.

10/17/13 Petrović. "Network/graph algorithms in algebraic statistics" , Joint discrete math - iCeNSa seminar, Notre Dame University

10/22/13 Fienberg. Panel Discussion: "Some Thoughts About Social Network Data And Tools for Their Analysis," SAMSI Network Research Workshop, for Program on Computational Methods in the Social Sciences, Research Triangle, NC.

11/07/13 Stasi. Manuscript 17. Lecture at Algebraic Statistics Seminar at IIT, Chicago.

11/13/13 Sadeghi. "Hierarchical models for independence structures of networks" poster at DARPA GRAPHS annual meeting. Washington DC.

11/13/13 Stasi. Manuscript 17. Poster at DARPA GRAPHS annual meeting. Washington DC.

11/12/13 Fienberg. "Developing Confidentiality and Privacy Protection Methods That Scale," International Year of Statistics Workshop on Future of Statistical Sciences, Royal Statistical Society, London, England.

11/14/13 Fienberg. Panel Presentation on large scale network data and challenges at Symposium on "Leveraging the Data Sciences," Part of the Presidential Inaugural Symposia Series, Crossing Boundaries, Transforming Lives, Carnegie Mellon University.

Other synergistic and organizational activities include the following:

Fienberg organized and ran a Statistical & Machine Learning Approaches to Network Experimentation Workshop at CMU April 21-23, 2013. The workshop brought together approximately 40 researchers to focus on how to design, implement and analyze experiments in networked environments, especially online ones. A key feature of many of the presentations was a form of network dependence that is related to our work on dk-graphs, which is ongoing research but directly related to the kinds of issues Fienberg points out in his review manuscript.

Petrović and Stasi invited and hosted the PIs from the Notre Dame and South Carolina Team in August and September 2013. Both Czabarka and Torockai gave colloquium talks in the Applied Math department at IIT and spent the day discussing research problems with the Team.

Petrović visited the Notre Dave team in October 2013, gave a talk at the seminar there (listed above), and discussed research problems of joint interest.

Stasi visited the Notre Dave team in November 2013, gave a talk at the seminar there, and discussed research problems of joint interest.

8. Deliverables

The set of products delivered during the first Phase *exceeds* the number of proposed deliverables. This is the result of the Team's research advancing toward the aims from later Phases, ahead of schedule.

A list of deliverables that the Team has made available follows, specifying the corresponding Tasks/deliverables in the project proposal.

I-1. Manuscripts **1** and **2** focus on deriving and studying the model polytopes for important ERGMs and exploits their properties to better characterize such models.

I-2.1. Manuscripts **8** and **10** motivate, explain and prove basic results toward the algebraic foundations for the toric ideals of hypergraphs.

I-3.1. Manuscript **3** characterizes, in geometric terms, the pseudo-likelihood, and compare the existence of MLE to that of pseudo-MLE.

III-2.2. Manuscript **17** also provides a computer program/ code used to carry out the goodness-of-fit analysis on real and simulated data for the p_1 random graph model.