Conditional Reliability and the Identification of Communities
Final Report

Charles J. Colbourn and Violet R. Syrotiuk
School of Computing, Informatics, and Decision Systems Engineering
Arizona State University

Final Report

This report outlines results on each of the six main tasks specified in the project. For convenience, the deliverables specified in the project are included *in slanted font* and the report is given following each deliverable. The deliverables are given here in the order most suited to the report, but numbered as in the original project.

In order to realize a general method to identify social communities residing among the peripheral end-user nodes of a network, the deliverables of the proposed research are:

3. *Develop a theory for conditional reliability. A report will be delivered that describes a model for the significance of the relationships* through*, rather than* to*, the core communities involved in network operation. The relationships will be inferred from common patterns of interaction with the highly interrelated core communities. The key result will be measures of the strengths of relationships among peripheral nodes when they are much more closely related to the core communities than to each other.*

and 4. *Develop a set of computational methods for conditional reliability. A report describing computational methods that identify weaker social interactions in the presence of very strong relationships in the physical network and strong ones in the logical network will be delivered.*

These have resulted in the development of reliability measures that have not been previously studied. Before commencing this project, we had updated a survey of current techniques for computation of network reliability [13]. In our application, however, it is necessary to determine that a community of nodes is interconnected. The basic reliability question is to determine the probability that a specified number of nodes is connected when some fixed nodes are to be included, and the remainder are to be chosen from a subset of nodes. Working with a PhD student, Toni Farley, we have unified a broad collection of measures of this type, and conducted a systematic literature review. A first version of the literature review is given in [17], and an abbreviated version appears in [20]. Because of the complexity of these computations, efficient algorithms for sparse networks have been devised, and are reported in [19].

Some details about the new measures follow: The k-terminal reliability measures are natural in network analysis, when one wants to know the probability that k specified nodes can communicate with each other in a given network. In network design, however, while it may be known that k-terminal operations will arise, it is unlikely that the identity of the specific nodes involved is known. These considerations motivated the definition of *(two-terminal) resilience* [5], the average two-terminal reliability over all choices $K \subseteq V$ with $|K| = 2$. This measures the expected ability of the network

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)* 11/9/09	**2. REPORT TYPE** Final Technical Report	**3. DATES COVERED** *(From - To)* June 2008 – Sept 2009

4. TITLE AND SUBTITLE

Conditional Reliability and the Identification of Communities

5a. CONTRACT NUMBER

5b. GRANT NUMBER
N00014-08-1-1069

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Charles J. Colbourn and Violet R. Syrotiuk

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Office for Research and
Sponsored Projects
Arizona State University
P.O. Box 872503
Tempe, AZ 85287-3503

8. PERFORMING ORGANIZATION REPORT NUMBER

CRS0146
08091205

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Office of Naval Research
875 North Randolph Street
Arlington, VA 22203-1995

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

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release; distribution is Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This document reports on results obtained in the investigation of a collection of methods for using indirect measurements of usage patterns in communication networks with the goal of identifying communities of users. The primary methods developed include a generalization of connectivity-based network reliability measures to multiterminal resilience measures that permit a more detailed classification of types of communication; techniques for identifying indirect connections based on patterns of common usage rather than direct communication; and (primarily) combinatorial techniques to support the determination of factors that can be used in determining relevant characteristics of such links in the identification of communities. In the latter vein, substantial progress on the construction of combinatorial arrays to permit factor characterization is outlined.

15. SUBJECT TERMS

Network reliability; conditional reliability; clustering; community identification

16. SECURITY CLASSIFICATION OF: U			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Charles J. Colbourn
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	6	**19b. TELEPHONE NUMBER** *(include area code)* 480-727-6631

to support a two-terminal operation, where the pair of terminals is selected uniformly at random. A generalization to k*-resilience* [18] instead yields the average k-terminal reliability over all choices $K \subseteq V$ with $|K| = k$.

Both k-terminal reliability and k-resilience address the question: What is the probability that k nodes can communicate? The difference is that for k-terminal reliability, the k communicating nodes are the k targets chosen in advance, while for k-resilience, the k are chosen uniformly at random. There is another natural way to interpret the question posed, when we are concerned with the existence of *any* k nodes in a connected component. We formalize the differences among these three interpretations by defining a common generalization of all three. Let $\mathcal{G} = (G, \mathbf{p}, \rho)$ with $G = (V, E)$ be a probabilistic graph. Let $H \subseteq V$, $|H| = h$; these are the *target nodes*. For a set $L \subseteq V$ and integer j, define $\Psi(G, L, j)$ to be 1 if G contains a connected component containing all vertices of L and at least j other vertices, 0 otherwise.

Define $\text{Con}((G = (V, E), \mathbf{p}, \mathbf{1}), H; h, i, j)$ to be

$$\frac{\sum_{\substack{I\subseteq V\setminus H\\ |I|=i}} \sum_{F\subseteq E} \left(\left(\prod_{e\in F} p_e \prod_{e\in E\setminus F}(1-p_e)\right) \Psi((V,F), H\cup I, j)\right)}{\binom{|V|-h}{i}}$$

An explanation in plain language is in order. When all nodes operate, this represents the expectation that $k = h + i + j$ nodes are connected, where the h nodes of H are selected in advance, the i nodes of I are chosen uniformly at random among the remaining nodes, and the j nodes can then be any remaining nodes.

Incorporating node failures in the definition is straightforward by considering the subgraph induced on the operational nodes. For a set $W \subseteq V$, define $E_W = \{e \in E : |e \cap W| = 2\}$; in other words, E_W contains the edges of the subgraph G_W induced on node set W. Then define

$$\begin{aligned}&\text{Con}((G = (V, E), \mathbf{p}, \rho), H; h, i, j) = \\ &\qquad \sum_{H\subseteq X\subseteq V} \left(\text{Con}((G_X, \mathbf{p}, \mathbf{1}), H; h, i, j) \prod_{x\in X} \rho_x \prod_{x\in V\setminus X}(1-\rho_x)\right)\end{aligned}$$

The definition at first seems somewhat unwieldy. Nevertheless, taking $h = k$ and $i = j = 0$, we obtain k-terminal reliability. Taking $i = k$ and $h = j = 0$, we obtain k-resilience. Taking $j = k$ and $h = i = 0$, we obtain the probability that $\mathcal{G}$ contains a component of size at least k, which we term the *kSet* problem. The definition also permits the analysis of more involved questions, such as determining the probability that h given nodes lie in a component of size at least k; or the probability that i nodes chosen uniformly at random lie in a component of size at least k. Such problems arise more frequently in reliability analysis than one might expect; see [17, 20].

6. *Develop clustering techniques to identify core communities. A report will be delivered that describes (1) the use of reliability computations in conjunction with clustering; and (2) the use of density-based approaches for the heuristic identification of clusters. In order to isolate communities at the periphery of the network, a necessary step is the determination of the "center" or "core" of the network. Clustering techniques based on density and on k-nearest-neighbour approaches will be compared to determine central clusters that are expected to form the physical backbone of the network.*

   The multiterminal resilience techniques developed in [19, 20] (described under Tasks 3 and 4 above) underlie a ranking of nodes. The easiest application is to determine the 'influence' of a node using its ability to communicate with each other node. We examined more sophisticated applications in [24]. Most relevant is the ability to rank nodes by their ability to connect communities of specified sizes to determine an *importance* for the node. Importance alone does not form a core cluster. The use of these measures provides a means for determining the importance of a node as a function of the reliability of each other node, that is determining the contribution of each node to the importance of each other. Computational tools have been developed to make these calculations for small networks.

5. *Assess the accuracy versus efficiency of reliability computations for transitive closure of relationships whose strengths are represented by probabilities. A report that describes the modelling of social networks using metrics based on network reliability will be delivered. This will investigate whether relationships are adequately captured by considering strengths of connections between pairs of nodes, or whether interactions among many entities are needed to capture social behaviour. It will also examine efficient bounding techniques for the relevant reliability measures to determine the feasibility of obtaining sufficiently accurate estimates of the strength of relationship for large networks.*

   Our work has focussed on citation networks. Under our direction, an Honors undergraduate student, David Weber, has collected data and formed large networks based on direct citation and indirect cocitation and bibliographic coupling data [24]. The analysis of these networks indicates that efficient bounding techniques based on edge-decompositions of the networks (in particular, the edge-disjoint pathset and cutset bounds and the consecutive pathset and cutset bounds referenced in [13]) are sufficiently accurate to distinguish among links and nodes. The inherent asymmetry of the networks involved appears to underlie the success of these methods. Limited application of the factoring method (see [13]) suffices in all cases examined to distinguish among groups of nodes with similar reliabilities. It is desirable still to adapt these bounds to the novel resilience measures discussed under Task 4. (This was begun by a Master's student, Kumaraguru Paramavisam, who left the program without completing the work.) In the interim, we have employed a crude Monte Carlo method.

   Modelling social networks based on their use of various communications networks relies on the development of appropriate link and node probabilities, which we discuss further under Tasks 1 and 2, below.

1. *Identify community usage patterns. A report identifying possible usage patterns that are characteristic of a community (i.e., common to the community but not frequently employed by other nodes) will be delivered. This includes the types of communications generated, the distributions of volumes and times of such communications, and the interaction with network services whose function is known.*

and 2. *Quantify the strength of the usage patterns. A report will be delivered that describes the relative merits of quantitative representations (particularly those based on sociometric and scientometric measures) for capturing inferred social relationships. The relationships derived will measure both social relationships and relationships reflecting network connection, which must be differentiated.*

   Our work has again focussed on citation networks, which represent a very simplified type of communication underlying a social network. These networks are severely limited by the types of communications measured, and hence investigation of internet data has been done. CAIDA [15] provides a repository of large-scale measurement data on various Internet functions; in addition, many public domain tools are available for analysis of these data [16]. The key concern is that as a result in the analysis of social structure the problem is 'data-rich' but 'information-poor'. The identification of possible usage patterns that are characteristic of a community do not appear to require data beyond that already collected through link, path, and network monitoring. Despite this, the differentiation between network and social communication, and among various types of social communication, appears to require a better initial characterization of link types and strengths in the communications network itself. Patterns of communication involve the types of traffic involved, the types of nodes involved, and the sequence and the timing of communication. Our results indicate that treating each of these characteristics independently is not sufficiently discriminating to separate physical communication from the logical and social communication that it supports. On a positive note, the indirect methods using cocitation analysis and bibliographic coupling do infer connections between actors that participate in similar types of communication; however, classifying the types of communication to determine this similarity is problematic.

Even in the citation data analyzed, it is understood that different actual citations serve different social purposes. Some are accepted as legitimate, such as those to well-known prior art, to key background (including relevant self-citation), or to related work cited to draw contrasts with the intended contribution. Others serve a different purpose unrelated to the contribution, for example to well-known figures in the field in an (often misguided) attempt to suggest the importance of the work, gratuitous self-citation, and attempts to manipulate the current ranking methods for impact. Our efforts to discriminate among these types of social communication have been severely constrained in two ways. First, the specific factors impacting the type and importance of a specific citation are unclear, and the number of factors to be considered is large. Secondly, the factors do not act in isolation from one another: Interactions among the factors are crucial. As a simple example from citation networks, the impacts of the author, author's institution, and the paper cited are correlated with the type of citation. But the impact of the journal, both at the time that the cited paper was published and at the time that the citing paper was published, correlate with the type of citation – and indeed the *change* in both rank and impact of the journal affect the type of the citation. In communications networks, we expect these problems to be more severe. Many more factors are present, and many more interactions should be anticipated.

An effective classification of links by type of communication is a prerequisite to the identification of communities with sufficient precision. Our research has therefore focussed on the determination and measurement of interactions among the many factors that are measured for physical links. Standard design-of-experiments techniques are inadequate in this context for a number of reasons. One is that many of the factors in network operation are measurable but not controllable. More importantly, before interactions can be measured, screening is needed to find the factors and interactions that may be relevant.

This problem arises in numerous different settings: screening using D-optimal designs [21], the location of interaction faults [12], approximate measurement using small sample spaces [1], internet tomography [4], and compressive sensing ([2], for example). These apparently different research areas all concern the identification and measurement among factors in which a signal or sample has a 'sparse' representation. Our research to date has concerned the unification, to the extent possible, of these different lines of investigation. We have established that the similarities among these topics are not just cosmetic, rather they arise from a deep connection in the underlying mathematics. Although this may appear to be a detour on the road to effective link characterization, we believe that understanding the fundamental similarities and differences among these is a necessary next step in finding a sufficiently accurate methodology to classify links, and in turn to use that link classification to isolate communities. In addition to providing the foundation for the specific problem in identifying social communities, this line of investigation can pay dividends in interaction fault location, signal processing, and sampling in sparse spaces generally.

In this line, we have completed a number of papers developing the foundations for the applications to factor location and screening. In [7, 8], powerful direct constructions using number theoretic methods have been developed. In [9], new recursive methods are developed. In [22, 23], a powerful computational search technique is developed. In [6, 14] a substantial generalization of 'perfect hash families' is introduced; these underlie worthwhile improvements in a column replacement strategy for combinatorial arrays; in [10, 11] an algebraic construction of such hash families is developed. These efforts are now being connected with compressive sensing, through the work of Colbourn's new PhD student, Chris McLean; and with error location through the work of Syrotiuk's new PhD student, Abraham Aldaco.

In the references to follow, research that is published or submitted and was funded partially or wholly under this project is marked with an asterisk ($\star$).

References

[1] Y. Azar, R. Motwani and J. Naor, Approximating arbitrary probability distributions using small sample spaces, *Combinatorica* 18 (1998), 151–171.

[2] R. Baraniuk, Compressive sensing, *IEEE Signal Processing Magazine* 24(4):118–121, 2007.

[3] M. O. Ball, C. J. Colbourn, and J. S. Provan. Network reliability. In M. O. Ball, T. L. Magnanti, C. L. Monma, and G. L. Nemhauser, editors, *Handbooks in Operations Research and Management Science: Network Models*, volume 7, chapter 11, pages 673–762. Elsevier Science B.V., Amsterdam, 1995.

[4] M. Coates, A. Hero, R. Nowak and B. Yu, Internet tomography, *IEEE Signal Processing Magazine*, 19 (2002), 47–65.

[5] C. J. Colbourn. Network resilience. *SIAM Journal on Algebraic and Discrete Methods*, 8(3):404–409, 1987.

[6] $\star$ C. J. Colbourn, Distributing hash families and covering arrays, *J. Combin. Inf. Syst. Sci.* (2009).

[7] $\star$ C. J. Colbourn, Covering arrays from cyclotomy, *Des. Codes Cryptogr.*, (2009).

[8] $\star$ C. J. Colbourn, G. Kéri, Covering arrays and existentially closed graphs, Lecture Notes in Computer Science 5557 (2009) 22–33.

[9] $\star$ C. J. Colbourn, G. Kéri, P. P. Rivas Soriano, J.-C. Schlage-Puchta, Covering and radius-covering arrays: Constructions and classification, *Discrete Applied Mathematics*, submitted for publication.

[10] $\star$ C. J. Colbourn and A. C. H. Ling, Linear hash families and forbidden configurations, *Des. Codes Cryptogr.* **59** (2009), 25–55.

[11] $\star$ C. J. Colbourn and A. C. H. Ling, A recursive construction for perfect hash families, *J. Math. Crypt.* (2009).

[12] C.J. Colbourn and D.W. McClary, Locating and detecting arrays for interaction faults, *Journal of Combinatorial Optimization* 15 (2008), 17–48.

[13] C.J. Colbourn and D.R. Shier, Computational issues in network reliability, *Encyclopedia of Statistics for Quality and Reliability*, F. Ruggeri, R.S. Kennet, F.W. Faltin (editors), Wiley, 2008.

[14] ⋆ C. J. Colbourn and J. Torres-Jiménez, Heterogeneous hash families and covering arrays, *Contemporary Mathematics*, submitted for publication.

[15] Cooperative Association for Internet Data Analysis, Data Collection at CAIDA – Research Topics, `http://www.caida.org/data/`, accessed 02 February 2009.

[16] L. Cottrell, Network monitoring tools, `http://www.slac.stanford.edu/xorg/nmtf/nmtf-tools.html`, accessed 03 April 2009.

[17] T. R. Farley, Multiterminal Reliability and Resilience, PhD thesis, Arizona State University, April 2009. (Advisor: C. J. Colbourn)

[18] T. R. Farley and C. J. Colbourn. Multiterminal resilience for series-parallel networks. *Networks*, 50(2):164–172, September 2007.

[19] ⋆ T. R. Farley and C. J. Colbourn, Multiterminal network connectedness on series-parallel networks, *Discrete Mathematics, Algorithms, and Applications* 1 (2009), 253-265.

[20] ⋆ T. R. Farley and C. J. Colbourn, Network reliability and resilience, *Proc. Design of Reliable Computer Networks (DRCN09)*, proceedings to appear.

[21] D.S. Hoskins, C.J. Colbourn, and M. Kulahci, Truncated D-Optimal Designs for Screening Experiments, *American Journal of Mathematical and Management Sciences*, to appear.

[22] ⋆ P. Nayeri, C. J. Colbourn, G. Konjevod, Randomized postoptimization of covering arrays, Lecture Notes in Computer Science 5874 (2009) 408–419.

[23] ⋆ P. Nayeri, C. J. Colbourn, G. Konjevod, Randomized postoptimization of covering arrays, *European Journal of Combinatorics*, submitted for publication.

[24] ⋆ D. Weber, Sociometric measures on networks, Undergraduate Honors Thesis, Arizona State University, April 2009. (Advisor: V. R. Syrotiuk)