The role of motifs in understanding behavior in social and engineered networks

Dave Braines[a], Diane Felmlee[b], Don Towsley[c], Kun Tu[c], Roger M. Whitaker[d], and Liam D. Turner[d]

[a]Emerging Technology, IBM United Kingdom Ltd, Hursley Park, Winchester, SO21 2JN, UK
[b]Dept. of Sociology & Criminology, Pennsylvania State University,
State College, PA 16802, USA
[c]College of Information & Computer Sciences, University of Massachusetts,
Amherst, MA 01003, USA
[d]School of Computer Science and Informatics, Cardiff University, Cardiff, CF24 3AA, UK

ABSTRACT

Within networks one can identify motifs that are significant recurring patterns of interaction between nodes. Here motifs are sub-graphs that occur more frequently than would be explained by random connections. Graphs can be used to model internal network structures of human groups, or links between groups, with group dynamics being governed by these structures. Graphs can also model behavior in engineered systems, and internal network structures can significantly affect dynamic behavior. A graph may only be partially visible (such as in hostile or coalition environments), however detectable network motifs may in some cases be reflective of the entire graph. We outline a research plan and describe basic network motifs and their properties, along with current analytic techniques for static and dynamic settings. We offer suggestions as to how network motif techniques can be applied to intra- or inter- group behavior, for example to detect whether multiple groups behave as a co-operative alliance, or whether coalition networks inter-operate in positive ways. As an example, we examine a complex time-series graph dataset relevant to coalition focused aspects of the class of networks under study, specifically related to the social network resulting from the authorship of academic papers within a coalition. We provide details of the basic analysis of this network over time and outline how this can be used as one of the datasets for our planned network motif research activities, especially with regards to the temporal and evolutionary aspects.

Keywords: Network motif, coalition, social network

1. INTRODUCTION

In recent years, research on social networks (i.e., connecting people) has expanded dramatically.[1] Studies repeatedly establish the importance of social network characteristics for a wide range of interaction processes, such as those in close relationships (e.g.,[2,3]), social organizations, science,[4] communication, crime and deviance, social media, as well as in war and terrorist activity (e.g.,[5,6]).

However, research seldom investigates sub-graphs within graphs that capture important human ties and interactions within these social networks. Such sub-graphs are referred to as motifs. Most studies of motifs have focused on their presence in static graphs and little is known about how they change over time. Changes are potentially useful in coalition environments, as changes in motif characteristics imply changes within the overall (hidden) network.

We are particularly focused on coalition networks where by coalition we mean a rapidly formed group of co-operating teams or organizations working towards a common objective or goal. Such a coalition may form rapidly and disperse after a short period, or may be more enduring in nature. We consider peace-keeping and disaster relief activities as good examples of such coalitions, where multiple organizations from military, civil

Further author information: Send correspondence to Dave Braines, dave_braines@uk.ibm.com

and NGO (Non-Governmental Organization) backgrounds are rapidly brought together into a specific operating environment and wish to quickly become productive by collaborating to progress their goals. Such coalitions are usually aligned in their desire to make progress against these goals in this environment even when they may not be in universal agreement about all factors.

In a coalition environment such as this, problems related to hostile and extreme external group-behavior may frequently emerge, such as in asymmetric warfare, insurgency and post conflict peace-keeping. A common understanding of how and why groups behave in specific ways is important for military intelligence, informing policy, resource deployment, and wider scenario modeling. However a persistent challenge concerns detecting and understanding the dynamics of groups that may only be partially visible. Group dynamics are governed by group internal network structure, i.e., connections (i.e., relationships) that allow a group to coordinate itself. Behavior, interactions and communication may only be visible between particular nodes in the network and at particular points in time, presenting a major obstacle for coherent modeling.

We are engaged in research to advance the state-of-the-art by exploring how network motifs can explain the external networks facing coalition operations, where noise and obfuscation is present. Network motifs refer to recurring, significant patterns of interaction between sets of nodes.[7,8] While an external network of interest may not be fully visible to the coalition, motifs represent important sub-graphs more likely to be visible, from which inferences can be made.

This paper outlines our research plans in this area for the Distributed Analytics and Information Science International Technology Alliance (DAIS ITA) and explores a particular time-series dataset identified for network motif analysis. Section 2 explores the technical approach to our planned research, along with key motivating questions, while Section 3 identifies the planned validation and experimentation. Section 4 explains the structure and statistics for the DAIS ITA Science Library, one of the key potential datasets for our research, along with initial results from basic motif analysis on a subset of the data. Section 5 concludes the paper.

2. RESEARCH CONTEXT AND PLANS

Motifs in networks are small sub-graphs in a graph that occur more frequently than can be explained by random occurrence. Network motifs can be applied to sub-graphs of a fixed size, e.g. dyads, representing a pair of nodes and the possible relationships between them; triads being the same but for three nodes and their possible relationships, and; tetrads being four node combinations. Links between nodes can be directed (one way) or undirected with different motif structures applying to undirected graphs.

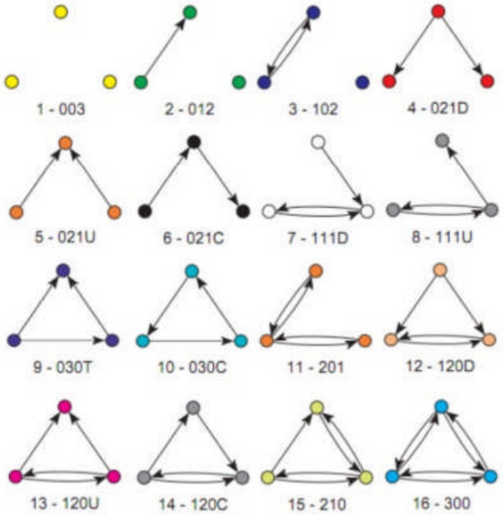


Figure 1. Examples of triadic motifs.[9]

For example, triads, or the (potential) ties connecting three actors, are considered to be the structural foundation of social networks.[10,11] Study of triads allow us to better understand a variety of network phenomena, including transitivity, the tendency for actor i to be tied to actor k if a tie exists between actor i and actor j and between actor j and actor k.

Our DAIS ITA research into the understanding of group behaviour, through these network motifs, is motivated by the following open research questions:

1. Do motifs characterize different groups of interest to DAIS ITA such as terrorist networks, different social networks, and communication ties? Do coalition networks exhibit different motifs, or are motifs universal across networks? What are good models for studying these networks?

2. How do motifs change over time in dynamic, mutable networks? Do different dynamic networks exhibit different temporal behavior? How should such behavior be modeled? Does a concept of temporal motif (T-motif) apply to dynamic change in networks?

3. Can motifs be used to gain insight into inter-group behavior, such as cooperative alliances or escalating conflict between external groups? Does the presence of motifs within a group provide insight into hidden structures and the mechanisms, such as the detection of hierarchies or informal coalescence, that drive in-group tension or cooperation?

4. Motifs appear in communication networks: do all coalition communication networks exhibit similar motifs? Does the interconnection of coalition networks exhibit motifs similar to those of the constituent networks?

5. What is the extent to which multipartite graphs (with multi-node, multi-edge, self loops and other similar features etc) can be used to elicit valuable network motif information? Does the presence of semantic information with the graphs (and therefore the ability to infer additional relationships) have a measurable or predictable effect on the network motif frequencies and distributions within these graphs?

6. Last, answers to the above questions rely on the availability of high-quality network datasets thus raising the question, how do we account for missing or erroneous data (presence of a link when one does not exist)?

We also intend to examine the presence of motifs in network data with the use of an exponential random graph model (ERGM).[12–14] The ERGM enables testing of nodal, dyadic, and structural tendencies (e.g.,[3]). Our goal is to examine the effects of key motifs, while controlling for other network properties in a multivariate framework.

In considering these six key questions we have outlined three main research threads with clearly identified goals. The first research thread focuses on *motifs in static networks*, the second on *motifs in dynamic networks*, and the third on the *role of motifs on inter-group behavior*. These three research tasks are detailed in the sub-sections that follow:

2.1 Comparisons of Metrics and Models of Multiple Social Networks

Goal: to determine presence and role of motifs in coalition-relevant human social networks from analysis of data.

Research on network motifs typically surround biological or physical networks, with human networks receiving little direct attention. The study of human networks represents a major, and crucial, gap in the literature that we intend to address.

We consider a computational approach that can be generically applied to many kinds of networks, and may be useful in identifying key network fragments that we expect to find in specific types of external groups. We will investigate whether identification of network motifs can be used to predict change over time in external groups. If we know that terrorist networks exhibit a tendency to be composed largely of triads that exhibit balance, or closure, for example, then we expect the presence of imbalanced triads to be unstable, and to change over time.

We plan to examine network motifs/sub-graphs, in social networks, and to undertake a comparison of differing genres of social interconnections (where "genres" represent different types of social networks, e.g. friendship, terrorist, twitter etc. See Section 3 for details). This includes development of algorithms for computing prevalence of different motifs in random networks. We will identify the important motifs associated with different genres of social networks by comparing their prevalence to that of random networks with the same degree sequence.[15] In preliminary work, we have begun to examine network motifs in several types of social networks, including those of multiple terrorist groups, email communication, friendship ties, twitter online messaging, travel routes, and advice ties. Initial results,[16] and earlier related research,[17] show that certain motifs appear to be universal in social networks. On the other hand, certain types of networks exhibit unique types of triadic motifs. In particular, aggressive, twitter communication networks are composed of a higher proportion of imbalanced, stressful triads than predicted, unlike most other types of social networks. Stressful triads are those which lack triadic closure and are more likely to be closed over time.

Motif compositions of terrorist groups are more similar to those among friends and positive alliance networks than to those of online communication or advice networks.[16] These results suggest that interconnections among insurgents are similar to those of friendship. These initial findings provide support for the argument that the terrorist groups studied here developed out of deep ties, as compared to the argument that such groups represented connections among relatively isolated cells. However, it is possible that new, online recruitment methods used to develop certain insurgencies may produce network ties that are more disparately connected than those observed so far.

So far we have only focused on social networks; we will also study motifs in communication networks, focusing on two aspects: 1) whether motifs are invariant across different coalition networks, and 2) how they compare with those in social networks.

In this particular paper we investigate the network structure of a social network dataset relating to co-authorship of academic publications within the DAIS ITA research community. Section 4 defines this network structure in some detail along with the data volumes for particular entities within the graph, and the evolution of the graph structure over time. This dataset will be one of many that we use to support the motif research defined above and it will serve as a useful contrast to the other existing datasets that have been explored in other publications. This dataset has coalition aspects (authors from numerous organizations) as well as social network links and a strong time series profile, showing the development of the publication and co-authorship data over time. We plan to identify the network motifs within various granularities of this dataset (i.e., creating different versions of the graph to support bipartite and tripartite complexities for example) and compare their frequency to random graphs as per the previous research, and also to identify whether these frequencies change in a predictable manner over time as the graph develops.

2.2 Dynamic Motifs in Dynamic Networks

Goal: to develop tools to analyze motifs in a temporal setting and to develop new insights into the latent behaviors of dynamic networks.

There has been little work on how motifs behave and change over time, with a few exceptions regarding the growth of sub-graphs.[18] Nevertheless, an examination of motifs over time has the potential to substantially further our understanding regarding the dynamic mutability of human groups.

We will develop new robust algorithms to study how sub-graphs change over time in a dynamic network. We will use these algorithms to characterize the temporal behavior of motifs and whether this behavior can be used to classify networks as well as to identify anomalous behavior. We will also attempt to extend the definition of motif to account for temporal changes with the goal of developing a definition for a "T-motif". Another goal is to develop a notation and/or language to define the motifs and capture their structure, dynamicity and tempo, and subsequently use this to predict wider network structures based on limited local observations, e.g. building on earlier work examining the evolution of triads over time.[9] Our ability to pursue these research directions require new representations of temporal motifs and new algorithms for studying their behavior. This is the focus of the research thread described here.

A dynamic network dataset consists of records that identify a contact between two individuals, either uni- or bi-directional, and a time stamp. This allows one to construct a sequence of directed (or undirected) graphs summarizing the dataset. We will explore two different approaches for studying the temporal behavior of sub-graphs/motifs in this setting. The first consists of performing a static motif analysis of each timestamp (as described previously) to create a motif summary of each snapshot. This could be the empirical sub-graph distribution for each snapshot or its entropy as examples. This produces a new time series that can be analyzed using classical techniques.

The above approach does not account for changes in sub-graphs associated with specific sets of nodes. The second approach will focus on summarizing changes in sub-graphs associated with the same nodes. This poses several challenges, the foremost being how to summarize these changes. We will explore the use of edit distance, i.e., the number of (directed) edge deletions and additions needed to transform a sub-graph in one snapshot to that in the next sub-graph. This can be used to produce different time series related to different edit distance statistics.

Another way of capturing changes is by calculating a transition probability matrix describing how sub-graphs change over time. Comparison of the stationary distribution of this Markov chain to the empirical sub-graph distribution will shed light on the role of randomness over time.

The problem has several dimensions. For example, the time granularity of individual snapshots can affect results. Moreover, behavioral differences as a function of snapshot granularity can provide useful information. There is the challenge of large datasets involving thousands to tens of thousands of nodes. We will adapt our recent results[19, 20] on sampling to analyze static graphs to the case of dynamic graphs. Last, we intend to explore different definitions of the novel concept of a temporal motif (T-motif) as part of this task. Last, datasets may be incomplete and/or replete with errors. We will model missing data as a consequence of a sampling process and extend our earlier work on sampling to handle it.

The development of the "T-motif" concept to describe the temporal change in network motifs over time is an important aspect of this work, resulting in succinct terminology and language for these T-motif profiles. We believe that the ability to describe the T-motif attributes of a particular graph will be valuable in supporting predictive analytics for the possible future development of the graph, or for inference to wider graph attributes when only a part of the graph is visible. We see this part of the research being well supported by the rich science library dataset described in Section 4.

2.3 Motifs and Emergence of Inter-group Behavior

Goal: to determine the extent to which motifs predict behavioral and structural features within and between groups.

In earlier research we explored event driven models for the evolution of group behavior.[21] This modeling approach involves actions taken in response to a social dilemma, based on individual and group-derived strategies. The social dilemma tests the extent of positivity or negativity towards a third party in the presence of interaction opportunities with others.

Actions that individuals take, both internally towards their in-group, and the response to an out-group, are influential to growth, cohesion and behavioral characteristics of the group. Motifs can capture interaction behavior within and between groups as a temporal sequence of events between actors. This provides opportunities for new insights: a bridge between individual actions and collective mission of a group that fuels conflict.

From a biological perspective, recent work[22] highlights the importance of motifs in the evolution of cooperation. In the DAIS ITA context of modeling group behavior, this can be significantly extended. Motifs can be tracked within the simulation of inter-group behavior to assess how a more diverse range of group behaviors, from cooperation and alliance, to hostility and warfare, emerges. Specifically we can use motifs to identify features that correlate with escalation and de-escalation of tension between groups in dynamic scenarios. Within groups, motifs are a basis for detecting potential mutation, such as breakaway sub-groups or opinion divergence. No framework exists for characterizing conditions that lead to internal division; motifs are ideal for supporting new insights into the substructures that lead to the escalation of divergence and tensions within groups.

We will adopt an agent-based simulation, which provides a dynamic context to observe the emergence of motifs. We will compare findings from simulations with real-world social media data that has been a-priori collected from other sources to identify whether the simulation findings relate to real online conflict situations. Additionally, we will explore offline conflict scenarios through open data provided by established ongoing projects (e.g. Social Conflict Analysis Database and Armed Conflict Location and Event Data Project, see Section 3 for further details regarding these potential datasets). This analysis aims to increase the value of information at coalition disposal, for example to better support situational understanding and decision-making.

Application of motifs within groups offers an important new mechanism to discover in-group hierarchy and structure in the presence of noise and obfuscation. These characteristics are typical in social networks belonging to subversive external groups that operate with restricted visibility. Motifs represent events between individuals, and snapshots of information flow. The importance of individual actors is reflected through their roles in multiple motifs, from which prediction of overall network structure and hierarchy is possible. Motif degree will be defined and examined in this context. This supports the identification of agents of influence: critical nodes in the network with enhanced roles in dissemination and connectivity of the group. Motifs are potentially well suited to this because they are not eliminated by partial network obfuscation. We will determine through simulation the extent to which detection of points of influence and structure is possible using motifs, in the presence of such obfuscation.

The dataset described in Section 4 is of particular relevance to the work described in this sub-task because we deeply understand the structure of the data and the network links for the dataset collected thus far. Using agent-based simulations to extend or mutate this dataset in ways that match credible real-world behaviours can yield a potentially large number of related synthetic datasets for motif analysis, especially to look for the motif characteristics described above. We are also able to simulate obfuscation or partial knowledge in a variety of forms, for example: removal of entire parts of the graph, random removal of nodes or edges to a predefined degree, or explicit removal of certain types of node or relationship. In each of these cases we will investigate the impact on the motif characteristics for the resulting degraded network and determine which factors most affect the motif distributions in each of these cases.

3. VALIDATION AND EXPERIMENTATION

As mentioned throughout this paper, our planned research will be validated through analysis of multiple datasets, seeking common approaches and patterns that can be shown to work across these datasets. In addition to this we plan for scenario simulation activities as well as the use of experimental data when possible. Details of each of these, in addition to a large set of candidate datasets are listed below. For this particular paper we are focused mainly on the "DAIS ITA Science Library" dataset (See Section 4 for full details), but some others are listed here to better outline the wider aspects of the research described earlier.

We will test our algorithms on a variety of datasets and use these datasets as a source of formal validation. For example: Terrorist network data sets;[23] Twitter Aggressive/Bullying data;[24] Friendship network data sets - National Longitudinal Study of Adolescent Health;[25] Networks of groups and actors involved in social or armed conflict;[26–28] Travel networks - US Airport networks;[29] and DAIS ITA networks - e.g., as recorded in the Science Library,[30] including the ability to have multiple time-series snapshots to show the evolution of key networks over time. It may also be possible to integrate additional relevant data from the earlier NIS ITA[31] in order to increase the available data volume and time period. See Section 4 for full details.

We will use simulations to emulate scenarios concerning the escalation of inter-group behavior, such as when cooperation is impeded by prejudice and hostility. We will also use simulation to develop test cases concerning the obfuscation of network structure. More details are given in Section 2.3.

Finally, we will test further our hypotheses using an experimental approach with the use of data from DAIS ITA Science Library. The key aspect here will be the generation of different experimental scenarios from the core underlying dataset.

4. EXAMPLE DATASET: SCIENCE LIBRARY

As described in Section 3, a key candidate dataset for the planned network motif research work is the DAIS ITA Science Library dataset, and this section gives a detailed description of that, along with key statistics and details of the time-series nature and growth of the graph over time.

The DAIS ITA Science Library is a publicly available website[30] that lists the publications of the DAIS ITA research program (with a second site providing the same experience for the earlier NIS ITA research program[31]). The data for the science library is held as a semantic graph using "ITA Controlled English" (CE),[32] providing an interesting and novel opportunity in this research since semantic graphs have not been a common focus in existing network motif analysis research. The semantic graph for the science library is comprised of instances (nodes) of given concepts (types or classes), connected together via named relationships (links), and with attribute values (properties). Much of the data in this graph is stated directly (in CE), but a substantial amount is inferred from logical inference rules (also written in CE). The resulting semantic knowledge graph is therefore a graph of nodes and links coming from the stated data and inferred data. From a network motif analysis perspective this is a directed multipartite graph. In fact the full graph is unwieldy in the current format and will be filtered down according to the key node types (and therefore links) to focus on the most relevant aspects from a social network and inter-relationship perspective. The data in this graph is made accessible via a browser based user interface that can be publicly accessed on the web.[30,31]

The key concepts and relationships are shown in Figure 2.

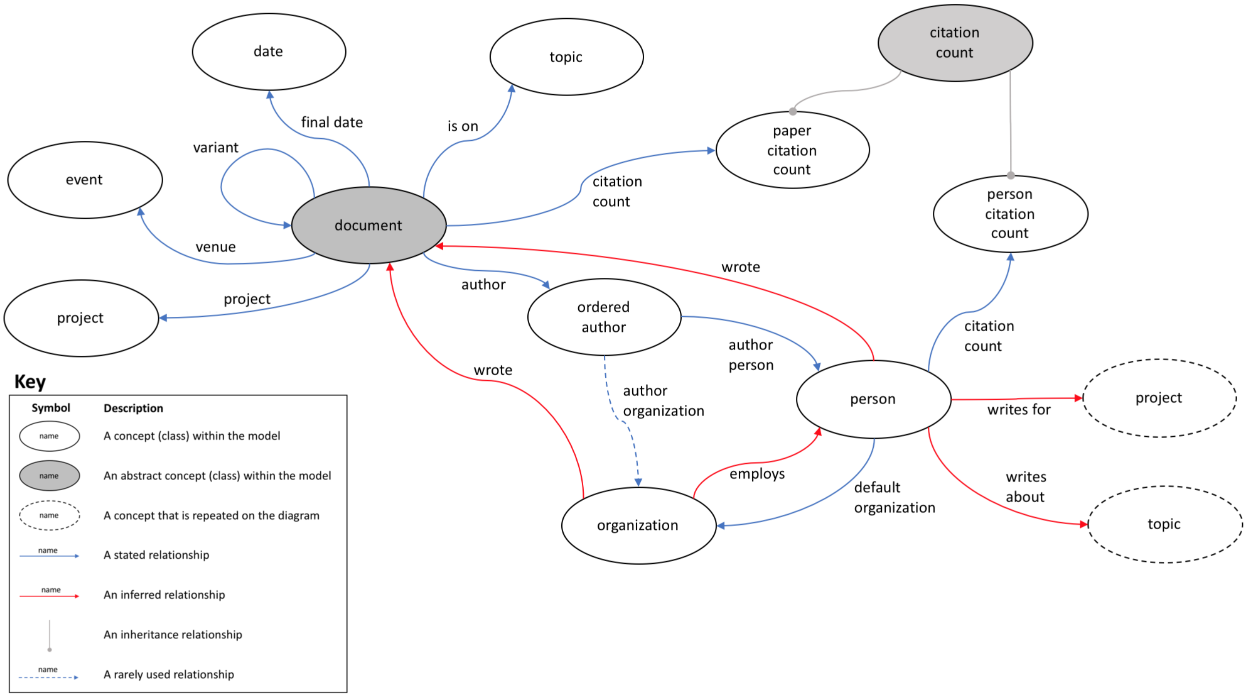


Figure 2. Science Library conceptual model.

The dataset in question is centered around the publication of academic papers and the social and topical networks that arise as a result. Given this publication-oriented focus for the dataset, we have chosen to identify the central concept within the graph as the *"Document"* concept. A *Document* corresponds to an academic publication created by the DAIS ITA research community and has links to *Authors* (and therefore *People*), and on to *Organizations*. *Documents*, *People* and *Organizations* are considered the fundamental concepts within the Science Library knowledge graph. A *Document* is also associated with *Projects*, *Topics*, *Events* and *Citation count* data and these are important secondary concepts. In addition to these there are many less important concepts and relationships within the graph, almost all of which will be ignored or removed as part of this research.

Stated more simply, the concepts (and therefore nodes and links) within the Science Library Graph break down into the following groups:

- **Central**:
  Document
- **Fundamental**:
  Document, Person (and therefore Author), Organization and Date
- **Important**:
  Project, Topic, Event, Citation Count

To help the reader understand the real-world meaning of this dataset a brief description of the central concept and each of the fundamental concepts is given below:

- **Document**
  A unique publication, such as a journal or conference paper.
- **Person**
  A unique individual who has authored one or more documents.
- **Author**
  The link between *Person* and *Document*. There will be one *Author* instance for every *Person* who is an author on every *Document*, enabling the unique *Documents* and *People* to be linked together without losing information.*
- **Organization**
  The unique organization that a *Person* claims affiliation to when authoring a *Document*.
- **Date**
  A unique month+year date which provides the time-series information within the graph, *e.g. Sep-2016*.
- **Project**
  A unique research project within the DAIS research program.
- **Topic**
  A unique topic identified as relevant to the published research.
- **Event**
  A unique event that is the publication venue for a *Document* (e.g. a conference or journal).
- **Citation count**
  A unique record for the citation data for *Documents* and *People*. This data is updated monthly giving one citation count instance per month per *Document* and per *Person* since the data started to be collected.

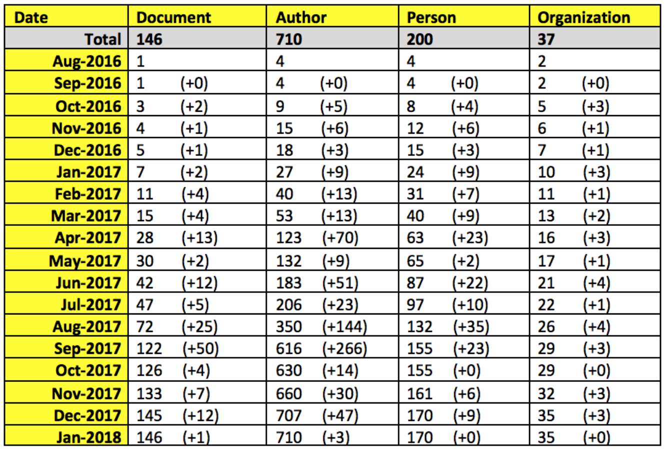
4.1 Time series data

In Section 2.2 we outlined the role of dynamic motifs in dynamic networks and introduced the concept of a "T-motif". The science library dataset has a strong time-series component, with the semantic graph evolving clearly across a well defined timeline that is instantiated as part of the graph. This is represented within the graph as the *Date* concept which is directly linked to the *Document* concept, our central concept. We can therefore add or remove documents based on their date, and this will affect all other nodes and links within the graph.

Documents have a "final date" link (along with numerous other date links), with final date being inferred by rules from the various other date links. This final date is at month granularity, allowing documents to be easily categorized into the month they were published.

Figure 3 shows the number of instances of each of the central and fundamental concepts within the graph. The month data shows how the time series evolves and the manner in which the numbers of unique instances of these concepts increase over time.

*In Figure 2 this is referred to as "Ordered Author" since each instance contains the position of the author in the overall author list for that *Document*.



Date	Document	Author	Person	Organization
Total	146	710	200	37
Aug-2016	1	4	4	2
Sep-2016	1 (+0)	4 (+0)	4 (+0)	2 (+0)
Oct-2016	3 (+2)	9 (+5)	8 (+4)	5 (+3)
Nov-2016	4 (+1)	15 (+6)	12 (+6)	6 (+1)
Dec-2016	5 (+1)	18 (+3)	15 (+3)	7 (+1)
Jan-2017	7 (+2)	27 (+9)	24 (+9)	10 (+3)
Feb-2017	11 (+4)	40 (+13)	31 (+7)	11 (+1)
Mar-2017	15 (+4)	53 (+13)	40 (+9)	13 (+2)
Apr-2017	28 (+13)	123 (+70)	63 (+23)	16 (+3)
May-2017	30 (+2)	132 (+9)	65 (+2)	17 (+1)
Jun-2017	42 (+12)	183 (+51)	87 (+22)	21 (+4)
Jul-2017	47 (+5)	206 (+23)	97 (+10)	22 (+1)
Aug-2017	72 (+25)	350 (+144)	132 (+35)	26 (+4)
Sep-2017	122 (+50)	616 (+266)	155 (+23)	29 (+3)
Oct-2017	126 (+4)	630 (+14)	155 (+0)	29 (+0)
Nov-2017	133 (+7)	660 (+30)	161 (+6)	32 (+3)
Dec-2017	145 (+12)	707 (+47)	170 (+9)	35 (+3)
Jan-2018	146 (+1)	710 (+3)	170 (+0)	35 (+0)

Figure 3. Time series evolution of science library graph.

Note that the total number of *People* defined in the graph is 200, whereas there are only 170 *People* who have authored a *Document* within the graph. The reason for this discrepancy is that at the beginning of the program, *Person* instances were created for every named researcher in the original consortium, but 30 of these *People* have not yet authored a *Document*. The same rationale explains the *Organization* total (37 defined, vs 35 that have published).

The graph will continue to be expanded as publications progress under the DAIS ITA research program, extending the month granularity time-series as explained above.

4.2 Overall graph statistics

The numbers in Figure 3 give details of the central and fundamental concepts and their instances. These numbers are quite small, however, and they should not be confused with the total number of nodes and links in the semantic graph. The overall statistics for the total graph are: **7043 nodes and 44,384 links**.

Figure 4 shows the central and fundamental concepts plotted against the time series, showing how the graph size and complexity develops over time. The substantial increases in particular months correlate to real-world events such as key conferences where multiple DAIS papers were published.

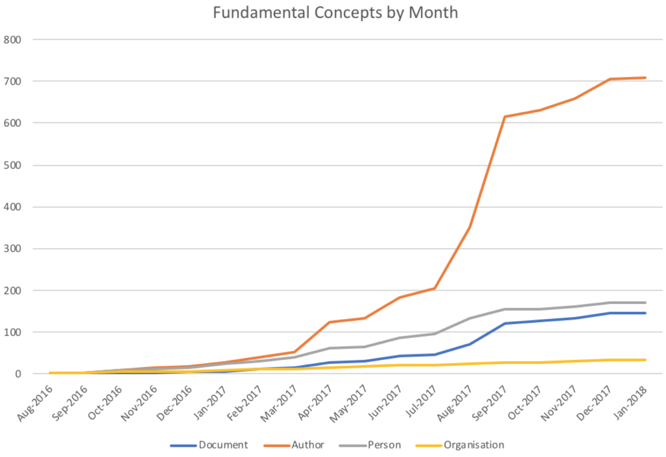


Figure 4. Growth of science library central and fundamental concepts by month.

Finally, we show in Figure 5 an example of the rendered simplified graph mid-way through the time series, showing the key relationships between the fundamental concepts in this network. The nodes are color coded according to their type, showing the multipartite nature of the graph, and the labels indicate some identifier for each of the nodes (e.g. author initials, paper ID or topic name). The size of each node indicates the number of links that node has to other nodes in the graph. This graph visualization helps to convey the full graph that will be the basis for creating further simplifications of the graph down to bipartite and tripartite variants to better enable the network motif analysis using existing tools and algorithms.

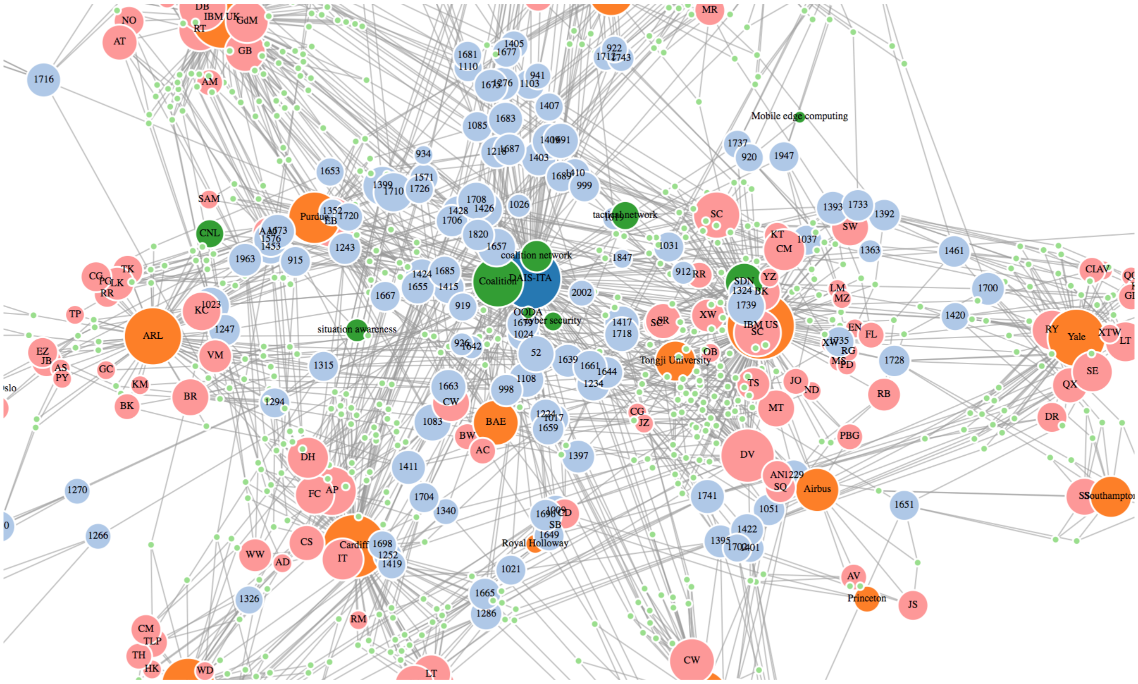


Figure 5. Example simplified graph for the central and fundamental concepts mid-way through the time series.

The data described above is publicly available under an Apache2 license, shared on GitHub[33] in the CE format mentioned previously.

4.3 Initial Subgraph Ratio Profile (SRP) analysis

Our analysis of the Science Library dataset so far is limited to specific subset of the graph against which we have run SRP (Subgraph Ratio Profile) analysis. The chosen subset of the data was deliberately limited to the single node type *Person* and the sole *co-author* relationship that exists between *Person* nodes. For this exercise we have 170 nodes (people) and 780 undirected links (*co-author* relationships) between these nodes.

Figure 6 shows SRP (without normalization) for undirected motifs (triads and tetrads whose nodes are connected) generated by random graphs with the same number of nodes and edges. In this result we note that motifs containing triangles have much higher frequency than random graphs.

The SRP for a triad t is computed with the following two equations:

$$\delta_t = \frac{N_t^{(G)} - \bar{N}_t^{(R)}}{N_t^{(G)} + \bar{N}_t^{(R)} + \epsilon} \tag{1}$$

$$SRP_t = \frac{\delta_t}{\sqrt{\sum(\delta_t)^2}} \tag{2}$$

where $N_t^{(G)}$ is the number of triad t observed in a network G, $\bar{N}_t^{(R)}$ is the average number of triad t appearing in a random network given the same degree-pair sequence as G. ϵ is a term to avoid dividing by zero.

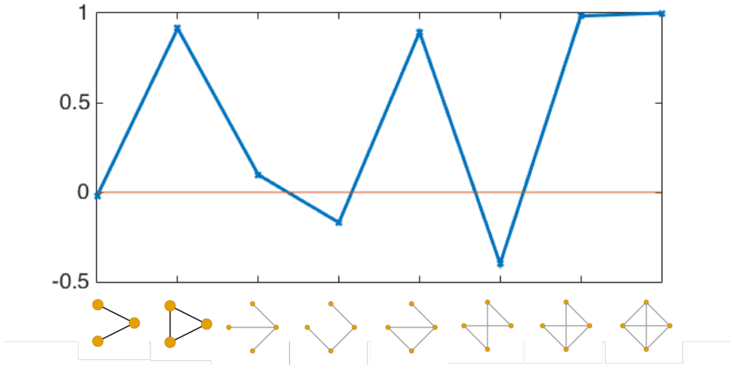


Figure 6. SRP for undirected motifs (triads and tetrads whose nodes are connected).

The prevalence of the triangle motifs in the data are likely to derive from conditions driving the behaviour of coalition members creating these publications. Co-authorship of publications (a single connection between two nodes) occurs in publications that involve multiple authors. In any research community there will be repeated publications between groups of authors, leading to well connected sub-graphs between these authors. On the DAIS ITA research program, from which this dataset is obtained, there is a specific focus on interdisciplinary collaboration and collaboration across organizational boundaries. For example by encouraging the formation of research teams comprised of UK and US members spanning academia, industry and government organizations. The composition of these teams fosters the creation of multi-author publications (such as this paper), building the co-author network across organizational and disciplinary boundaries. Furthermore, with key researchers engaged in multiple research teams this again develops the co-authorship network.

Further analysis will take into account more node types as well as a wider variety of relationships in addition to performing the analyses at different points in the time series sequence.

5. CONCLUSION AND NEXT STEPS

This paper has outlined the network motif research plans for the DAIS ITA research program and has described in some detail the "Science Library" dataset as one of the candidate datasets to be used as an experimental basis for some of the proposed research. A basic analysis of the dataset was given, along with some statistical information and an outline of the time-series evolution of the semantic graph that constitutes that data set. We plan to progress through the research plan and develop the various permutations of this dataset and others to support our research tasks.

The Science Library dataset is manifested as a semantic knowledge graph and as such it is straightforward to create modified permutations of the graph by extending (or suppressing) the existing semantic attributes and thereby inferring more (or less) links within the graph. This capability raises a number of questions, mainly as sub-questions to the multipartite research question (Q5) outlined earlier in Section 2, which we would like to pursue during the phase of dataset generation and selection, for example:

1. Do the types of nodes selected have a predictable effect on the motif distributions in the resulting graph?
2. Do inferred (computed) relationships modify the motif distributions when compared to the raw underlying graph? And if so, do those changes in motif distributions have any predictive capabilities?

Answers to the above questions will help inform our strategy for generating the dataset permutations to support our research, and may yield interesting insights in their own right, especially when taking into account any such measurable changes over time for these networks.

ACKNOWLEDGMENTS

This research was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence under Agreement Numbers W911NF-06-3-0001 and W911NF-16-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

REFERENCES

[1] Felmlee, D. and Sinclair, C. H., [*Social Networks and Personal Relationships*], Cambridge University Press (2018, forthcoming).

[2] Faris, R. and Felmlee, D., "Casualties of social combat: School networks of peer victimization and their consequences," *American Sociological Review* **79**(2), 228–257 (2014).

[3] Felmlee, D. and Faris, R., "Toxic ties: networks of friendship, dating, and cyber victimization," *Social psychology quarterly* **79**(3), 243–262 (2016).

[4] Lungeanu, A. and Contractor, N. S., "The effects of diversity and network ties on innovations: The emergence of a new scientific field," *American Behavioral Scientist* **59**(5), 548–564 (2015).

[5] Everton, S. F., [*Disrupting dark networks*], vol. 34, Cambridge University Press (2012).

[6] Krebs, V. E., "Mapping networks of terrorist cells," *Connections* **24**(3), 43–52 (2002).

[7] Alon, U., "Network motifs: theory and experimental approaches," *Nature Reviews Genetics* **8**(6), 450 (2007).

[8] Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U., "Network motifs in the transcriptional regulation network of escherichia coli," *Nature genetics* **31**(1), 64 (2002).

[9] Doroud, M., Bhattacharyya, P., Wu, S. F., and Felmlee, D., "The evolution of ego-centric triads: A microscopic approach toward predicting macroscopic network properties," in [*Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*], 172–179, IEEE (2011).

[10] Holland, P. W. and Leinhardt, S., "The statistical analysis of local structure in social networks," (1974).

[11] Wasserman, S. and Faust, K., [*Social network analysis: Methods and applications*], vol. 8, Cambridge university press (1994).

[12] Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., and Morris, M., "ergm: A package to fit, simulate and diagnose exponential-family models for networks," *Journal of statistical software* **24**(3), nihpa54860 (2008).

[13] Morris, M., Handcock, M. S., and Hunter, D. R., "Specification of exponential-family random graph models: terms and computational aspects," *Journal of statistical software* **24**(4), 1548 (2008).

[14] Wasserman, S. and Pattison, P., "Logit models and logistic regressions for social networks: I. an introduction to markov graphs andp," *Psychometrika* **61**(3), 401–425 (1996).

[15] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U., "Network motifs: simple building blocks of complex networks," *Science* **298**(5594), 824–827 (2002).

[16] Felmlee, D., McMillan, C., Towsley, D., and Whitaker, R., "Social network motifs: A comparison of building blocks across multiple social networks," Annual Meetings of the American Sociological Association, Philadelphia, US (2018).

[17] Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M., and Alon, U., "Superfamilies of evolved and designed networks," *Science* **303**(5663), 1538–1542 (2004).

[18] Paranjape, A., Benson, A. R., and Leskovec, J., "Motifs in temporal networks," in [*Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*], 601–610, ACM (2017).

[19] Wang, P., Lui, J. C., Towsley, D., and Zhao, J., "Minfer: A method of inferring motif statistics from sampled edges," in [*Data Engineering (ICDE), 2016 IEEE 32nd International Conference on*], 1050–1061, IEEE (2016).

[20] Wang, P., Qi, Y., Lui, J. C., Towsley, D., Zhao, J., and Tao, J., "Inferring higher-order structure statistics of large networks from sampled edges," *IEEE Transactions on Knowledge and Data Engineering* **99** (2017).

[21] Whitaker, R. M., Turner, L., Colombo, G., Verma, D., Felmlee, D., and Pearson, G., "Intra-group tension under inter-group conflict: a generative model using group social norms and identity," in [*International Conference on Applied Human Factors and Ergonomics*], 167–179, Springer (2017).

[22] Gianetto, D. A. and Heydari, B., "Sparse cliques trump scale-free networks in coordination and competition," *Scientific reports* **6**, 21870 (2016).

[23] "John jay and artis transnational terrorism database (JJATT)." http://doitapps.jjay.cuny.edu/jjatt/data.php (2009). Accessed: 2018-03-06.

[24] Sterner, G. and Felmlee, D., "The social networks of cyberbullying on twitter," *International Journal of Technoethics (IJT)* **8**(2), 1–15 (2017).

[25] Harris, K. M., "The national longitudinal study of adolescent health: Research design," *http://www.cpc.unc.edu/projects/addhealth/design* **1** (2011).

[26] "Social conflict analysis database (SCAD)." https://www.strausscenter.org/scad.html (2012). Accessed: 2018-03-06.

[27] "Social conflict analysis database (SCAD)." https://www.strausscenter.org/o.html (2012). Accessed: 2018-03-06.

[28] "Armed conflict location and event data project (ACLED)." https://www.acleddata.com/data/ (2015). Accessed: 2018-03-06.

[29] "Airport, airline and route data." https://openflights.org/data.html (2017). Accessed: 2018-03-06.

[30] "Science library: publications of the distributed analytics and information science international technology alliance (DAIS ITA)." http://sl.dais-ita.org/science-library (2016). Accessed: 2018-03-06.

[31] "Science library: publications of the network and information science international technology alliance (NIS ITA)." http://nis-ita.org/science-library (2017). Accessed: 2018-03-06.

[32] Braines, D., Mott, D., Laws, S., de Mel, G., and Pham, T., "Controlled english to facilitate human/machine analytical processing," in [*Next-Generation Analyst*], **8758**, 875808, International Society for Optics and Photonics (2013).

[33] "Science library data for DAIS ITA." https://github.com/ce-store/sl-data-dais (2017). Accessed: 2018-03-06.