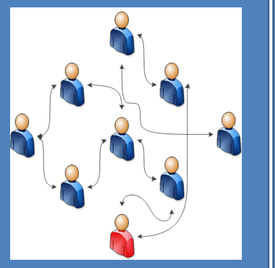
Now You See Us – Now You Don't: Community Detection in Multilevel Networks

Eric McGinnis, Ian McCulloh, Christine Sowa, Janis Butkevics, Alexander Perrone
The Johns Hopkins University Applied Physics Lab and Army Research Lab

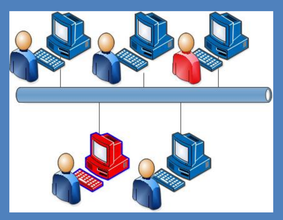
Abstract
Individuals today maintain personas in a multitude of social networks ranging from online social media networks to offline collaboration networks and multiple networks in between. These networks are hardly disjoint and exhibit significant overlap, where a relationship in one network may increase the likelihood and dynamics of a relationship in another network. In some social networks, relationships are readily apparent, where in other social networks they remain hidden. Multilevel network data was collected on twelve different faculty departments of a small university in North America. This includes publicly available data such as co-authorship and website co-occurrence as well as private data such as file sharing and email communications. Community detection approaches are applied to each network and commonalities are investigated. Recommendations are provided based on applications including defense against social engineering attacks and studies of organizational behavior.

Motivation

As social and computer networks become more complex, they play an increasingly important role in people's lives and day-to-day organizational operations. Security-conscious organizations have realized that a robust, multidimensional approach may be required for both defensive operations to address an ever-increasing variety of threats.



Malicious actors may have contact with and/or influence other individuals in a network



Insider Threats and Compromised Hosts May Present Anomalous Signatures on a Network

Classic approaches including antivirus, firewalls, auditing, security awareness training, and access controls may be insufficient to counter the risks posed by Insider Threats, victims of Targeted Social Engineering, and Advanced Persistent Threats. A holistic model could characterize users, services, and hosts to facilitate automatic detection and countering of anomalous activities.

Social Theory

"A man may be regarded as the point of intersection of an indefinite number of circles representing social groups, having as many arcs passing through him as there are groups." – Cooley, 1902

"The structure of a network is dependent upon the constraint and size of the underlying foci. Highly constraining foci will create close-knit clusters of various sizes depending upon the size of the foci." – Feld, 1981

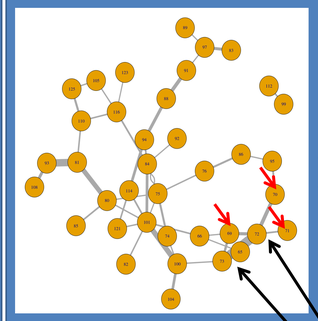
Research Objectives

- Identify methods to extract social networks from available data
- Investigate the correlation/overlap between social networks
- Investigate pattern of life differences based on social grouping
- Explore the position/role of central actors in the social network

Data Sources

A variety of publicly available and closed, novel data sources were used to support this work, including:

- Publicly available data from a social networking site obtained via web scraping
- Publicly available data from a research institution's staff profile pages including biographies and paper authorship
- Internal network data including email records and file sharing activity



Internal communications for the largest department at the University

IT Staff Lead – direct reports are marked in red

IT Director

Collaboration via Email

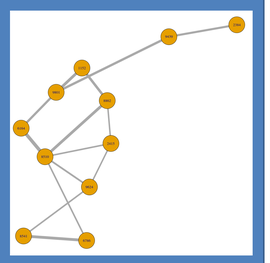
Email is problematic due to mass distribution lists that do not involve meaningful interaction, differences in user email behavior, and individual motivation for email use.

We remove emails with over 150 recipients (Dunbar rule). The sum-inverse-recipient-count (SIRC) weights links by the level of interpersonal relation.

Blockmodels and Reduced Graphs

Blockmodels partition the adjacency matrix by department affiliation. The density of the group by group SIRC allows estimation of inter-departmental collaboration.

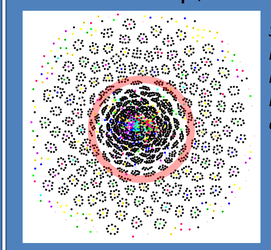
- Significant collaboration between Engineering, Physics, and Math Departments (shared research centers)
- Significant collaboration between Social Sciences, English, and History (shared research centers)
- Math and Engineering Departments are key points in the network



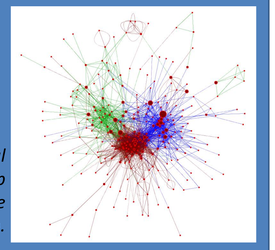
Links weighted by normalized interdepartmental email frequency

Social Network and Website Data

Surprisingly, there is little overlap between social media, email, co-authorship, and web page similarity.



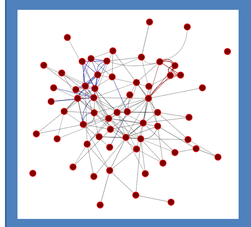
Social Network users generate links to related profiles through browsing activities. Core represents major research centers.



Networks developed using social network data, co-authorship data, and web profile cosine similarity show little overlap.

E-mail networks

Email tends to fulfill communication requirements outside of social media behavior. Co-authorship tends to occur among more senior faculty that have been in the department for a longer time. The social media behavior node cluster includes faculty more engaged in the publicity of their research.

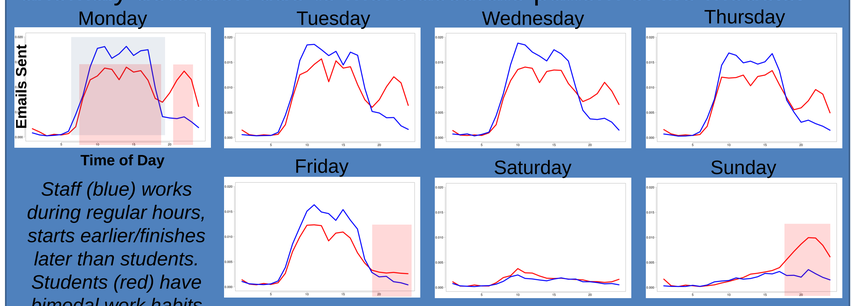


Network on the left represents the largest department's email (gray), co-authorship (blue), and social media behavior (red). The network on the right omits email to show the contrast.

Requirements for writing, publishing, and indexing delay the availability of co-authorship data

Pattern-of-Life Analysis via Email

Social network analysis allows identification of clusters of actors with different network behavior. This can allow more powerful anomaly detection due to more accurate pattern of life models.



Staff (blue) works during regular hours, starts earlier/finishes later than students. Students (red) have bimodal work habits

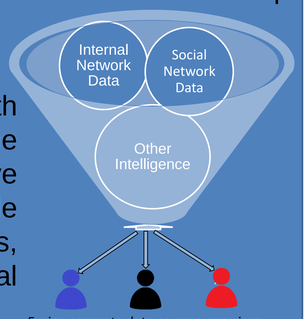
Little student work on Friday nights…significant work Sunday nights

Conclusions

Our work shows progress in exploring both publicly available and internal data. We have proposed methods to produce network maps, identify key players in networks, and visualize patterns in user activity. Social network community detection allows more accurate pattern-of-life models. The lack of overlap between network relations is surprising.

Future Work

Combining our novel data sources with robust persona matching explored in the literature may yield significant defensive opportunities. These opportunities include the identification of vulnerable users, systems, and files as determined by social network measures.



Fusing separate data sources promises improved threat identification

Acknowledgements

This is a project of the Johns Hopkins University Applied Physics Laboratory. This work was supported in part by the U.S. Army Research Laboratory.