November 7, 2011

Untangling Arachne's Web: Utilizing Social Network Science, AUTOMAP, and ORA to Analyze Social Change in Tanzania

Francis Nimick and John Ringquist PhD

Network Science Center, West Point
www.netscience.usma.edu
845.938.0804

Untangling Arachne's Web: Utilizing Social Network Science, AUTOMAP, and ORA to Analyze Social Change in Tanzania

ABSTRACT

The analysis of historical networks by traditional methods of historical inquiry does not lend itself well to research that seeks to determine the influence of past historical actors in a quantifiable, verifiable manner. However, historical actors can be placed within a network context with the use of social network analysis software. There are multiple options for network generation, some direct and easily employed through spreadsheet applications; other methods of analysis such as historical text analysis require significant amounts of time but create deeply detailed networks with many sub-contexts. We have chosen to employ two methods that utilize software developed by Carnegie Mellon University: ORA and AUTOMAP. These programs give the capacity to generate networks manually (ORA), or through a text mining application such as AUTOMAP.

Background

When the topic of determining the viability of frontier markets in Tanzania was first broached for research, a team of historians attempted to answer the question: What was the hidden factor that gave stability to the Tanzanian state between 1970 and 1994, a state with a single party and yet no coups or praetorian competition for power within the various groups that composed the state apparatus. The hidden factor, it appeared, was a close civil-military relationship that

continued after market liberalization and political reform. We then determined that it would be useful to determine points in the relationship between the military and the state during which social networks developed, matured, and then fractured into new networks. Conceptualizing networks as resilient, yet set into a social network matrix that was almost crystalline in structure gave us a set of perimeters that we chose to test, starting with the training of the military forces of the Tanzanian People's Defense Force (TPDF). The Tanganyika Rifles had mutinied in 1964, and the TPDF was a new hybrid force that contained both Chama Cha Mapinduzi (CCM) political commissars and TPDF officers in its organizational structure. Training, we hypothesized, was where the two networks would overlap. Before we could test the hypothesis, we needed to determine the network analysis method. The first method, automated mapping, was tested prior to our decision to utilize ORA for our first network analysis, the diplomatic records of TPDF and CCM officers training in foreign locales between 1976 and 1977.

AUTOMATED NETWORK MAPPING[1]

Automated network mapping involves the generation of networks from raw plain text. This is a time-consuming process that requires dedicated researchers with the time and patience to construct search perimeters. The networks generated can be either simple semantic networks

[1] PORTIONS OF THIS SECTION ARE FROM FRANCIS NIMICK, "MANUAL & AUTOMATED TEXT MINING & NETWORK GENERATION: A BRIEF ANALYSIS," UNPUBLISHED ANALYSIS PAPER FOR THE UNITED STATES MILITARY NETWORK SCIENCE CENTER, AUGUST 24, 2011.

that link words based on frequency and proximity, or detailed meta-networks that classify items based on a user-created thesaurus. Although frequency analysis has merit, especially for large articles or material downloaded from the internet, meta-analysis is a more precise search method for specific concepts linked to a pre-determined set of perimeters. Search perimeters are vital for AUTOMAP; in short, words matter because they are the basis for text analysis.

Despite the sheer amount of time necessary to utilize AUTOMAP, automated network mapping has several advantages over manual entry methods: consistent networks in which the same automated mapping procedure will produce identical networks when applied to identical documents in identical contexts. Therefore AUTOMAP would lend itself readily to analysis of diplomatic records or monthly reports such as those that would be generated on a regular basis utilizing a common format. These consistent results, because they are not as readily subject to human bias (as with manual entry), provide a solid framework for statistical analysis. In addition, automated mapping scales very well in theory, as the related documents we would be working with contain similar documents and terms. After the initial document has been analyzed and the initial thesauri generated, any additional related documents can then be analyzed with minimal additional time expenditure. In aggregate this translates to the ability to analyze many similar documents after the thesaurus was constructed. The advantages are obvious, but so are the drawbacks, especially for projects that intend to encompass multiple dimensions and thousands of documents. The thesaurus for such an endeavor would be almost unimaginably large and complex.

Despite the many advantages of automated network mapping, it is still not a viable option for large projects such as ours, one in which we would attempt to analyze multiple factors in a longitudinal study. The output of a text analysis is highly dependent upon the methods employed to analyze it; absolute consistency is essential to derive reliable results. Consistency therefore constrains the work perimeters of teams, and while not necessary onerous for a large group or researchers, is a huge obstacle to small groups. When only one person can work on a document or set of documents at a time; and if the documents are to be analyzed as a single network, the time cost of verifying consistency and terminology increases exponentially with document length. A simple test run with a single, thematically uncomplicated journal article of less than thirty pages resulted in a thesaurus of over 150 terms. Research assistant Francis Nimick estimated that at a document with a length of a hundred pages would require a thesaurus that was so large and complicated that the time required for consistency verification would render network analysis unfeasible all practical intents for teams without access to more advanced forms of computers for analysis.

Text analysis software suffers from other shortcomings as well: AUTOMAP had difficulty analyzing complex and high-level text. AUTOMAP was incapable of distinguishing between words that refer to different objects, i.e. last names in context, thus necessitating manual user intervention by modifying the source text. Unfortunately this shortcoming creates a situation in which users cannot assume that any given analysis is automatically valid; text must

be edited to identify errors and redundancies. The advantages of scalability must be matched against the need to check every analysis for distortions, a drawback that led our research team to consider ORA as the primary research method. However, no analysis system is perfect and we needed to weigh the advantages of ORA before proceeding with data set analysis.

Manual network mapping has several advantages over automated network mapping. Manual network mapping does not require the user to extract or process text, thus eliminating a large portion of the time required for automated mapping. Manual mapping also eliminates the multiple steps of preprocessing and manual verification involved in automated mapping and thesaurus creation. Manual network mapping also produces networks that are more relevant and accurate than those produced by automatic mapping. Manual mapping ensures that the only concepts represented in the network are those that are relevant to the research at hand; automated network mapping depends upon the simple textual relation of words, where manual mapping is capable of including a deeper analysis of the underlying large-scale concepts. Where automated mapping commonly generates networks that assign high importance to common terms such as "not" or "which," manually generated networks include only important pieces of information. In addition, manual network mapping eliminates the tedious step of creating meta-network thesauri. The assignment of categories to terms can be done on a case-by-case basis during the process of network generation. Despite the many advantages of ORA-based network analysis, there are some shortcomings of significance.

Manual network mapping is prone to issues of consistency when more than one user constructs networks. Because ORA may be utilized to apply value or "weight" to links in a network, it is essential that all members of a research team agree on the weight assignment perimeters. If formal definitions are not determined, and arbitrary values are assigned that vary by user, the resultant networks cannot be relied upon to accurately represent any quantifiable aspect of the source document. Link weights have bearing on network structures and give meaning to node placement and relations with other nodes. Centrality is contingent on node weight: were we to eliminate link weights entirely and create a binary network, we would be dogged by the analogous question of how strong of a relation between variables must exist in the sources before formal inclusion in a network? Manual network mapping introduces bias through determination of node weight value. Therefore, a hybrid technique may bear the best results.

Francis Nimick suggested the utilization of a semi-automated network based on a two-step process. During the first step all relevant nodes in the document would be created, with categories assigned to them. Second, link weights would be determined by a controlled procedure. Relation strengths were assigned as follows: No relation – link weight 0; knowledge of – link weight 10; opposition to – link weight 20; support of – link weight 30. The implication of linked or common network relationships in these weights is clear. Much as one could expect indifference from strangers (link weight 0), neutrality from associates (10); hostility from enemies or rivals (20) or close allies (30), so would link weights reflect relationship strength.

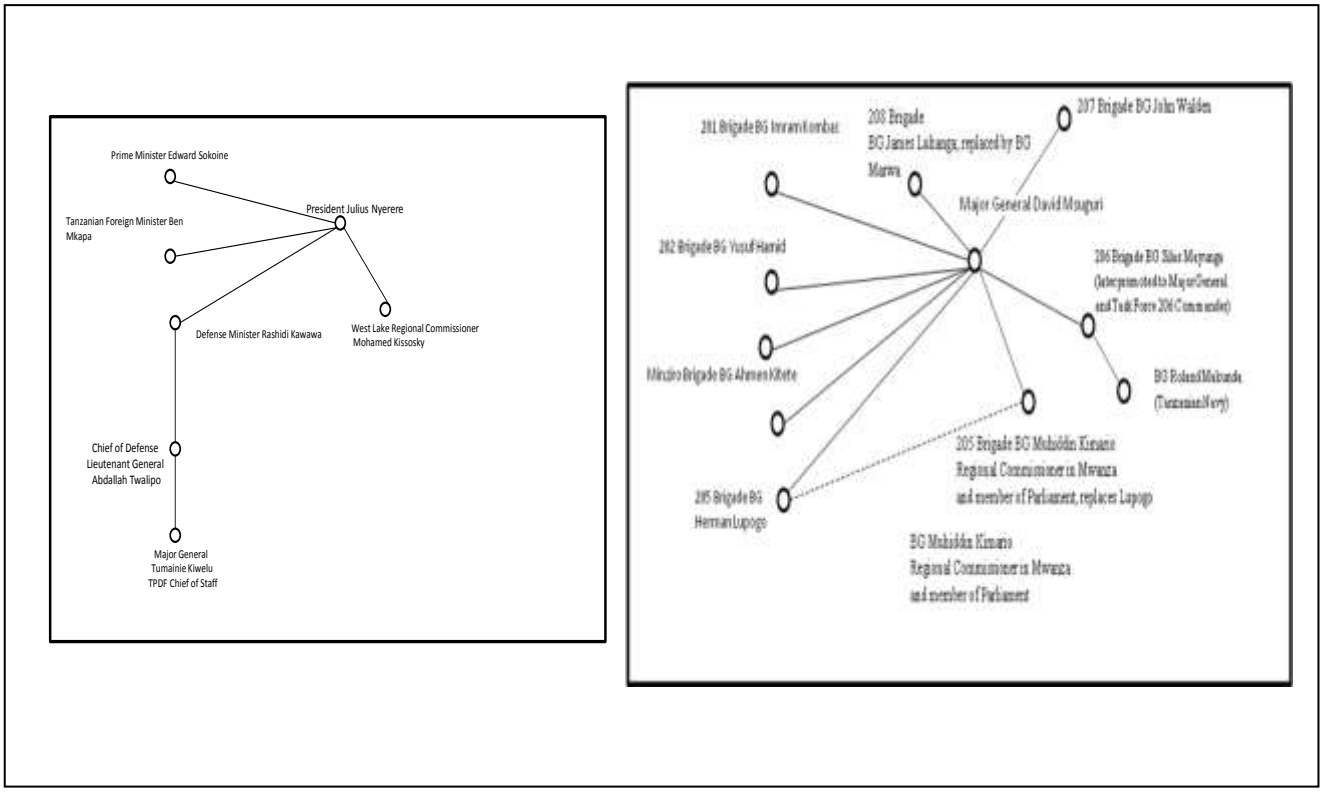
The creation and application of weight is especially compelling for historical documents such as personal correspondence or political diatribes. Weight allocation also allows for variance in network resiliency and political gaming: for example, if person A supports person B on one topic and opposes him on another, the link weight between them would be approximately 15. While imperfect, this method seems to be the best for analyzing large quantities of text such as those involved in our project.

Application of Carnegie Mellon's ORA to Construct Complex Networks

After weighing the merits of the two differing analytical methods, we chose to create a case study in which we would analyze the training variable using a data set of 130 different training documents from the TPDF archives in Lugalo, Tanzania. ORA, as Carnegie Mellon's web site describes it, is "an analytic toolbox that can handle multi-mode, multi-plex, multi-level networks. It can identify key players, groups and vulnerabilities, model network changes over time, and perform COA analysis."[2] Our analytical purposes did not encompass nearly all the abilities of ORA, and network construction defaulted to analyses of 3-4 variables in order to keep networks uncluttered, while allowing for testing the impact of un-weighted variables upon network construction. As a further test of our networks we constructed a social network based on what information we could derive from secondary sources on the war between Tanzania and Uganda. Traditional methods of historical analysis could draw out some network relations and make

[2] "ORA," www.casos.cs.cmu.edu/projects/ora

assumptions based on secondary sources or first-level relationships in documents, but the assembly of data into graphical representations appeared weak at best.



Network Diagram 1[3]

Archival sources demand analysis to derive meaning. However what meaning is derived is dependent on the researcher and their own innate knowledge of the material. Therefore two people could have widely varying analyses of the same material. Network Diagram 1 could be dismissed as expressing a logical military rank-derived network with little hidden import. Rank determines strength in a network and stability could be attributed to the influence of military discipline. Recent research in the Tanzanian national archives revealed a network of junior

[3] Tony Avirgan and Martha Honey, *War in Uganda: The Legacy of Idi Amin* (Westport, Connecticut: L. Hill, 1982).

officer cadre that clustered around their educational training and political indoctrination. After analyzing diplomatic requests and state visits, the career trajectory of these particular officers assumed interesting trajectories. The networks that developed could prove decisive in determining if our hypothesis held merit.

Individuals skilled in network analysis study a series of variables such as educational training possessed the same homophily, to determine if a combination of similar and dissimilar factors, offers *substructures of foci* centered on relationships "formed in workplaces, voluntary associations, and neighborhoods."[4] Homophily analysis offers a method by which to analyze the configurations that develop based on education, assignment location, and political association between the CCM and TPDF. In the case of the TPDF training network homophily could be linked to training location or duty assignment as easily as it could be linked to factors such as ethnicity or religion. Individual military branch training could replace ethnic identity as infantrymen associated with like individuals in similar locales at the same time.

A second tool to determine the relative degree of influence exercise by individuals within a network is to assess the leaders that are assigned to the group (node) by the CCM, or who are linked to the CCM through assignment or locale. A leader's "insularity" is one way to determine the most connected node in a network, and the connectedness of leaders to one another.[5] Soldiers that serve together should logically be expected to know one another and if united in

[4] David R. Schaefer, "A Configurational Approach to Homophily Using lattice Visualization," *Connections*, Vol 30, No. 2, (December, 2010): 22-23.
[5] Samuel Arbesman and Nicholas A. Christakis, "Leadership Insularity: A New Measure of Connectivity Between Central Nodes in Networks," *Connections*, Vol. 30, No. 1 (April, 2010): 4-10.

common cause, be expected to maintain relationships after departing the same locale for new duty assignments. The influence of individuals in networks could be measured by ORA through algorithmic analysis and a "nodal influence" value calculated for particularly well connected nodes.

A third tool that could be utilized to determine connectivity, stability, and influence within the TPDF and CCM social network is positional dynamics analysis as an exchange of power in return for loyalty, friendship, or favors. Individuals that exercise power in networks have multiple connections to nodes and we are able to determine the amount of power by their centrality and location in a strong network. [6] Networks can and do often rearrange themselves when a node, or central connection is broken, but resilient social networks that exercise influence and have deep connections to other networks will have multiple connections across several common linkages. When a central figure is removed from a position for whatever reason, the network may shift to the second most connected individual. Quite simply, when people are reassigned, demoted, promoted, or no longer serve in an active network, they integrate with a new network. However, seldom are all ties severed in a social network.

Network Diagram 1 is an example of a network that may be developed strictly along the lines of participants in the 1978-1979 Uganda War is the command network of top officers that played

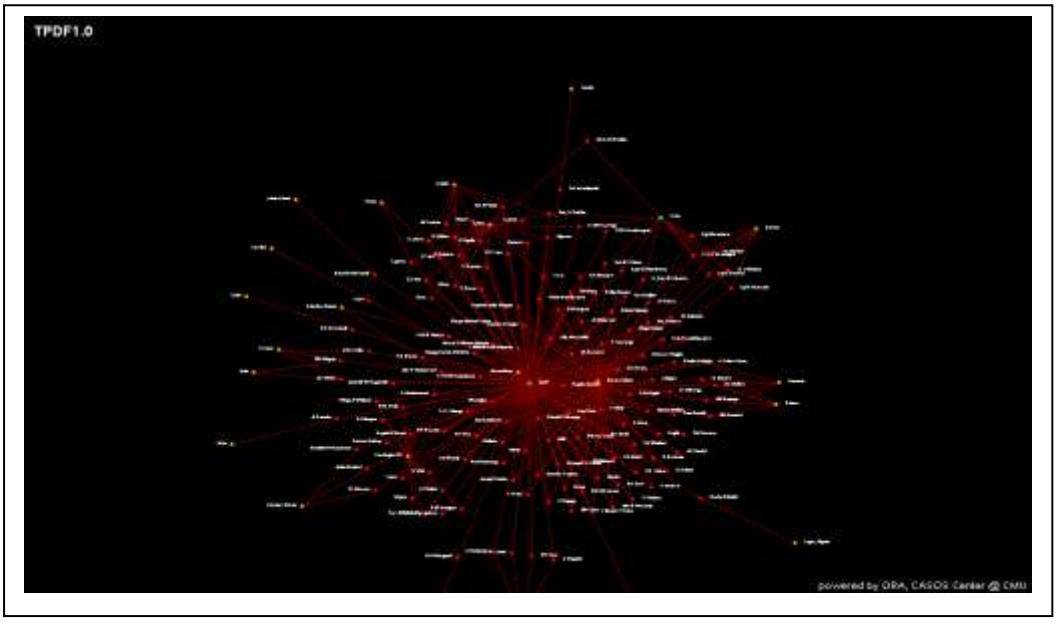
[6] David Easley and Jon Kleinberg, *Networks, Crowds and Markets: Reasoning About a Highly Connected World* (Cambridge: Cambridge University Press, 2010), 302. Individuals that exercise a high degree of power will be expected to occupy a central position in a network, much like an imaginary dot in the center of an X combines all incoming legs of the letter into a center unifying position. The strong position and multiple connections imply that the person, or node, is in a position to dispense favors or in the event of external threats to the network, dispense aid and monitor loyalty (connectedness).

major roles in the counterattack into Uganda.[7] Arranging officers into clusters reveals a civil-military network, but little can be drawn from the network except for what is reasonable; the top ranking military individual would have the most positional power due to rank and not social network influence. There is no weighted value except what is assumed based on historical knowledge, and no mathematically defined statistics accompany the model. It is sterile and lacks value for predictive analysis. The military network (on right) reveals tight lines of control that express a formal chain of command. The reality is that while Major General David Msuguri could be expected to have the greatest influence, he was not as influential as Lieutenant Colonel Ben Msuya of Brigadier General Silas Mayunga's 206 Brigade in the final phases of Kampala's liberation from Idi Amin's forces. Indivudual influence may appear out of scale in a standard archive-derived network, but a social network analysis of TPDF training experience in 1976-1977 reveals an example of a remarkably well-linked individual, a key node in branch of the main network.

Lieutenant Colonel Cathles was a key link to training assignments in China and Angola, but more importantly, he also served as a go-between for the CCM and TPDF for at least seven other individuals in the network diagram. Few other persons in the TPDF-CCM training network diagram had this degree of impact on junior officers. Most junior officers linked with other junior officers and few touched senior officers in the network. Not surprisingly, Lieutenant

[7] Tony Avirgan and Martha Honey, *War in Uganda: The Legacy of Idi Amin* (Westport, Connecticut: L. Hill, 1982), 79. Major General David Msuguri is an exceptional person in this group of leaders because his position was determined by his personal connection with President Nyerere. Although Msuguri was a former King's African Rifles soldier who served in WWII, he "Would have been chief defense forces except that he happened to be from Nyerere's home village, Butiama, and the Tanzanian president didn't want to be accused of favoritism."

Colonel Cathles was the TPDF training officer and could be expected to have had a significant role in junior officers' careers. Such a position could have exercised great power and influence. Yet, Cathles is not mentioned in the secondary sources as playing a significant role in the Ugandan War. Why? He was in China during the war.

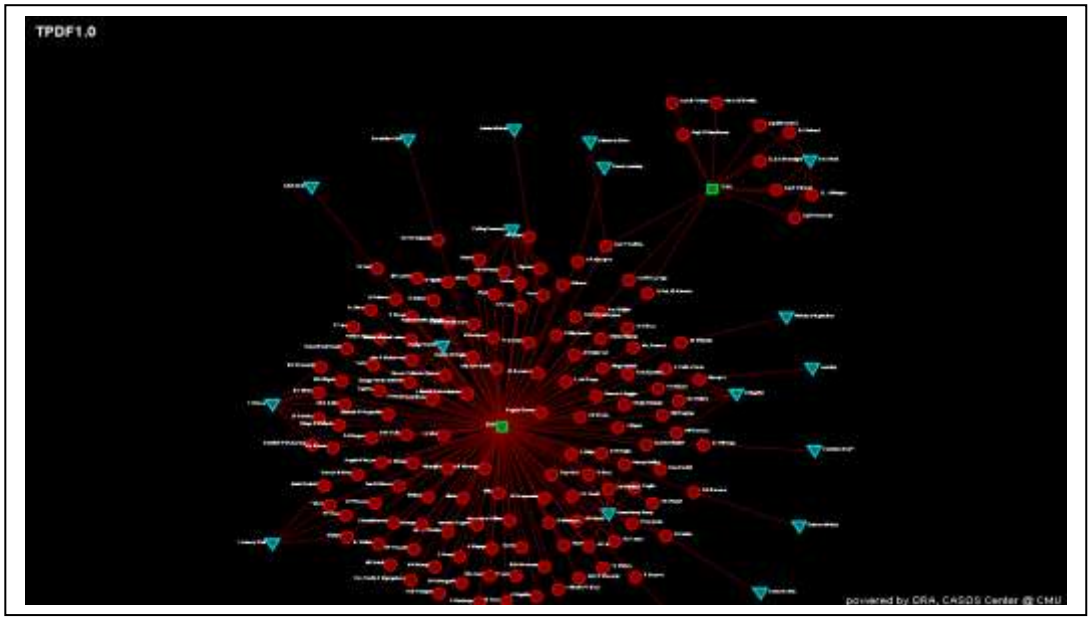


TPDF/CCM Social Network 1

Word Cloud Representating Training and Actors

Although a word cloud revealed little about any one individual in the network based on frequency analysis, a social network diagram isolated Cathles' impact on the social network and the connections between the various actors in such a way that Cathles' relationship was unmistakable.

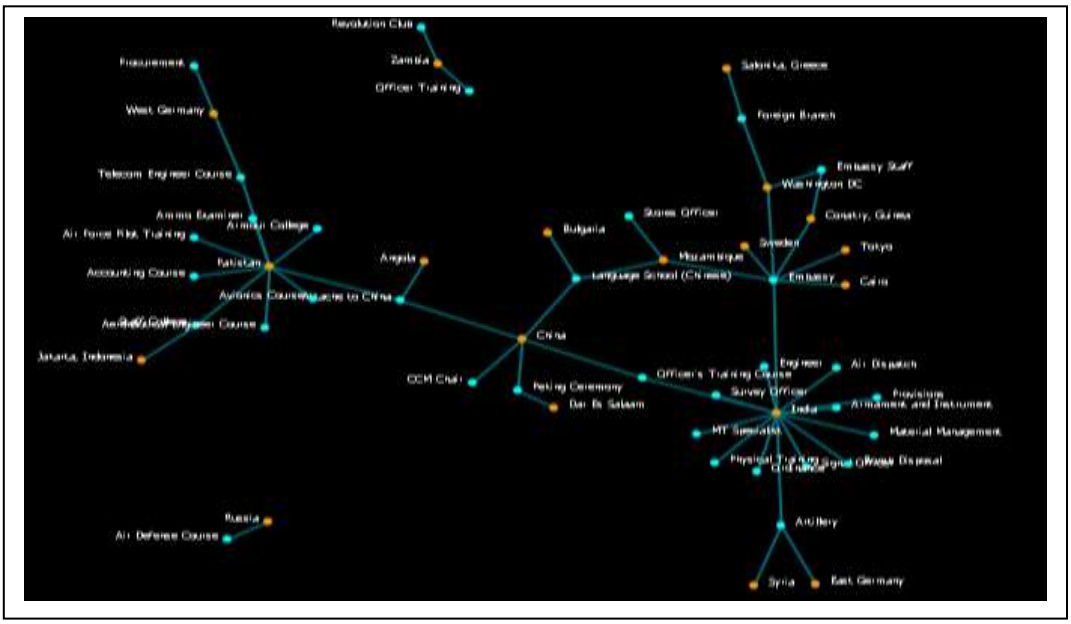
Furthermore, the relationship between the CCM and the TPDF was quickly demonstrated as junior officers clustered around the political party, represented on the left of the diagram. The



TPDF/CCM Social Network 2

relationship was expected given the heavy political indoctrination of junior officers following the 1964 coup. The degree to which the CCM was involvement in military affairs, and vice versa,

indicates that the state contained a deeply integrated civil-military infrastructure from which senior officers were isolated by duty assignments. Senior officers could not exercise influence in a 1976-1977 TPDF network because of their lack of significant interaction in the social network.



TPDF/CCM Social Network 3: Training Locations

Training locations appear to be significant in the construction of social networks, as was hypothesized. Among networks there were some more important than others. Embassies, regional security offices, and political party offices all played central roles. Military training was therefore contingent on political influence, and not just military requirements. Training was an indirect measure for exercising influence and patronage.

Conclusions

Although our research is still ongoing, the fact that social network analysis has revealed stratification in the TPDF ranks and a tight allegiance to CCM-generated alliances indicates that a longitudinal study is merited. Our data set is a small portion of what archival information awaits researchers. However, if we develop a AUTOMAP thesaurus that will allow us to complete analysis of the remaining .5 linear feet of diplomatic documents, it can be hypothesized that the resulting social network will be a dense cloud of tightly linked individuals that we can then track into our next two analytical nodes, 1985 and 1994. These two years coincide with events of great political import: structural adjustment and political change from single to multi-party democracy.

The utility of social network analysis for researchers of past societies is clear. The work to operationalize data of this sort for contemporary operations through analysis of contemporary open-source data is also compelling. However, the dual constraints of time and consistency mitigate against the employment of this technique to derive results in a short suspense situation. Conservatively any researcher curious about the application of ORA or AUTOMAP to a social or historical question should anticipate significant personnel investment and the use of multiple computers for dedicated analysis.

Works Cited

Arbesman, Samuel and Nicholas A. Christakis. "Leadership Insularity: A new Measure of Connectivity Between Central Nodes in Networks." *Connections*, Vol. 30, No. 1 (April 2010):4-10.

Avirgan Tony and Martha Honey. *War in Uganda: The Legacy of Idi Amin*. Westport, Connecticut: L. Hill, 1982.

Easley, David and Jon Kleinberg. *Networks, Crowds and Markets: Reasoning About a Highly Connected World*. Cambridge: Cambridge University Press, 2010.

NIMICK, FRANCIS "MANUAL & AUTOMATED TEXT MINING & NETWORK GENERATION: A BRIEF ANALYSIS," UNPUBLISHED ANALYSIS PAPER FOR THE UNITED STATES MILITARY NETWORK SCIENCE CENTER, AUGUST 24, 2011.

"ORA," www.casos.cs.cmu.edu/projects/ora

Schaefer, David R. "A Configurational Approach to Homophily Using Lattice Visualization." *Connections*, Vol 30, No. 2, (December, 2010): 21-40.