Data Fusion for Regionally Aligned Forces

David Beskow, Caitlin Rowe, Jed Lee, Maxat Nugmanov, and Ruben Vargas
Department of Systems Engineering
United States Military Academy
West Point, NY 10996
Email: david.beskow@usma.edu

***Abstract*—Our research uses Network Centrality metrics applied to the Global Knowledge Graph (GKG) to provide military tactical commanders with a tool that identifies influential individuals at the national or sub-national level. GKG aggregates the people, organizations, locations, and themes of English news sources from across the world using an enhanced TABARI (Text Analysis by Augmented Replacement Instructions) algorithm. We use several programming languages to download, query, and analyze the GKG data daily. We then apply network centrality metrics to identify the most important people and organizations in a selected area of interest. The most influential names are merged with a two-sentence description from Wikipedia. These algorithms were configured into an open-source web application. This model and tool will allow tactical leaders of regionally aligned units to rapidly understand and engage influential persons in their operating environment.**

I. Introduction

As the war in Afghanistan is nearing its end, the U.S. Army is focusing on becoming a more agile force. By definition, an agile force must be able to deploy rapidly and integrate into a multiplicity of operational environments. The U.S. Army recently created the Regionally Alligned Forces (RAF) organizational construct in order to facilitate agility and engagement. Forces are regionally aligned in order to better connect the Army with partner nation's forces and build people-to-people relationships. It also reduces the number of environments that these forces must be prepared to operate in. In addition, the U.S. Army recently added a warfighting function called "engagement" to its doctrine in order to train soldiers to effectively work with host nations and regional partners. However, as soldiers engage with people in various regions around the world, the need for relevant and valuable information at the sub national level is increasing. Open source data can provide military leaders with significant intelligence if analyzed and visualized effectively.

In order to understand the importance of fusing data to provide tactical leaders with open source information, it is important to understand the context in which this information will be implemented. For example, as a result of the recent escalation of the Ebola virus outbreak in 2014, the U.S. Army deployed 3,200 soldiers in October 2014 to Monrovia, Liberia as the Joint Force Command for Operation United Assistance. Their mission was to support interagency humanitarian efforts and supervise the construction of Ebola treatment units. These soldiers were deployed to Liberia under the command of U.S. Army Africa [1]. These Soldiers operate in an unfamiliar and austere environment where a virus has killed over 3,300 people and continues to spread. It is important that military leaders in these situations understand their operating environment, particularly influential actors. These actors include government officials, tribal/village leaders, local celebrities, and key social media actors that each have local influence. In order to influence an area toward a desired state, commanders must work through those persons and organizations that have influence. Open source data contains a wealth of information that can help military leaders understand these complex environments and identify who has influence in a region [1].

A. Background

Since March 2013, the U.S. Army has used the Regionally Aligned Forces (RAF) concept to posture and orient Brigade Combat Teams across the world. RAF is an extension of the "U.S. Army's vision for providing Combatant Commanders with tailored, responsive and consistently available Army forces" [2] and serves to not only optimize the use of resources, but to connect the Army on a global, strategic scale by engaging with partner nation forces and building people-to-people relationships. Some functions of RAF include providing crisis response, operational support, and theater security cooperation in various areas. These forces serve around the world in support of the Unified Combatant Commands (Joint Commands provide command and control for all U.S. forces operating in specific Areas of Responsibility, i.e. Africa Command). These Regionally Aligned Forces deploy into austere environments at a rapidly increasing frequency [2].

Success for these Regionally Aligend Forces hinges on their ability to effectively engage the leaders and influencers at the national and sub-national level. The U.S. Army has several Warfighting Functions that it defines as essential elements of combat power. These functions include *mission command*, *movement/maneuver*, *intelligence*, *fires*, *sustainment*, and *protection*. Recently, the U.S. Army Training and Doctrine Command (TRADOC) added its seventh warfighting function: *engagement*. Lieutenant General Keith C. Walker describes the engagement warfighting function as "the capabilities and skills necessary to work with host nations, regional partners, and indigenous populations in a culturally attuned manner that allows bridging language barriers, opening lines of communication and connections with key political and military leaders in a way that is both immediate and lasting" [3]. Increasing military leader's understanding of their environment and the influencers within it will naturally increase his/her effectiveness in engagement.

Currently, the intelligence community employs five primary methods to collect intelligence for various government departments. The five methods are human intelligence (HUMINT), signals intelligence (SIGINT), imagery intelligence (IMINT), measurement and signatures intelligence (MASINT), and open-source intelligence (OSINT). While there are various methods of effectively collecting information, our work will focus on OSINT. Open-source data is easily accessible and can provide a vast amounts of information. The challenge in the OSINT community, however, is filtering, extracting, and presenting information that is valuable and actionable. Our tool aims to fuse open-source data quickly to identify influential individuals and infrastructure at sub-national levels in order to help tactical leaders of regionally aligned units understand and engage their areas of interest before and during temporary deployments to austere environments [4].

B. Data

Our models consume open source data provided by the Global Knowledge Graph (GKG). GKG is part of the Global Database of Events, Language, and Tone (GDELT) that was created and is curated by Kalev Leeraru [5]. GDELT is an initiative to construct a catalog of human societal-scale behavior and beliefs across all countries from 1979 to present. The data is created using enhanced Textual Analysis By Augmented Replacement Instructions (TABARI) coding of numerous English language news and report sources. GDELT is geo-referenced and distinguishes between ethnic and religious affiliations of various state and non-state actors. Additionally, GDELT is able to examine and classify emotion-based indicators. This dataset is relatively large ($\approx 350GB$ in a MySQL database), and therefore requires appropriate analytic tools to query and analyze. Our use of GDELT is one of the first times that a DoD entity has used this new and large data set for a direct application. The Global Knowledge Graph (GKG), derived from GDELT, "connects the world's people, organizations, locations, themes, counts, and emotions into a single holistic network over the entire planet" [5]. Each row of GKG Data consists of the events, people, and places that are found in an article, as well as the tone (or sentiment) of the article. Our research used the *people* field in the GKG data to draw connections between individuals, and the *location* field to subset the data by geography. An example of the data is given in Table I.

Our research assumes that individuals co-mentioned in an article are in some way connected. We represent these connections between individuals (and some organizations and infrastructure) in a network graph with vertices and undirected edges. Since many of these connections are not social connections, we do not characterize this network as a social network. Applying network centrality measures to this network allows us to measure and rank influencers. The *location* field in the GKG data allowed us to extract geographically based subsets of the data and apply these network centrality measures to city, province, national, or regional networks.

We decided to fuse Wikipedia data with the GKG data in order to provide commanders and staff a brief description of each of the influencers in a region. A list of unknown names is of limited value to military leaders, forcing them to conduct significant background research to decide which individuals are worth additional study. By adding a two sentence description of individuals taken from Wikipedia, we can provide immediate summary information allowing a staff to quickly determine which individuals to research further and present to the commander.

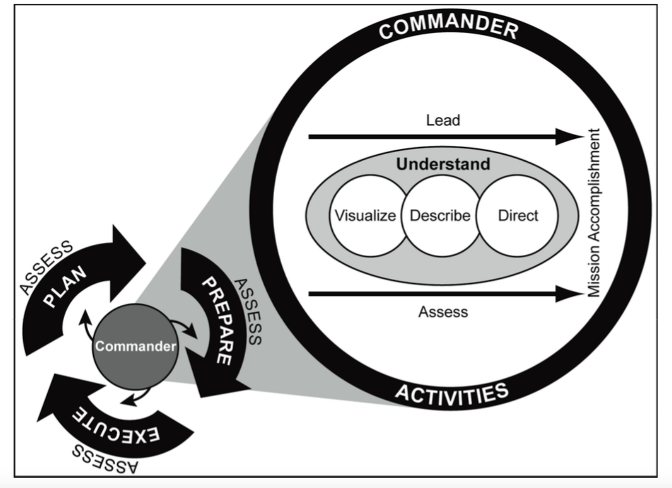

Fig. 1. The Army Operations Process [6]

II. Literature Review

A. Military Doctrine

The Army's framework for executing leadership is the *operations process*. Army doctrine states that

> Commanders, supported by their staffs, use the operations process to drive the conceptual and detailed planning necessary to understand, visualize, and describe their operational environment; make and articulate decisions; and direct, lead, and assess military operations. [6]

This process is depicted visually in Figure 1. This process clearly demonstrates the need for the commander (and his/her staff) to *visualize* the environment and the problems/challenges that it contains. This process of visualization enables the commander to apply the right solution to the right problem. *Visualizing* an environment and problem set is both an art and a science, and requires a commander and staff to build and maintain situational understanding. Our research and decision support tool endeavours to provide the science part of situational understanding and support the commander in visualizing challenging problem sets.

B. Network Centrality

While graph theory has its origins dating all the way back to Leonhard Euler in the 18th Century, its application in human networks was delayed until the mid 20th century. With computational improvements over the last fifty years as well the development of the internet, network science rapidly grew to model the internet itself as well as the people that connect over it. Today, with large computers and vast amounts of data at our fingertips, network science uses graph theory, social theory, communications research, and data science to assist in characterizing and analyzing various types of networks [7].

TABLE I. EXAMPLE OF GLOBAL KNOWLEDGE GRAPH DATA (TAKEN FROM BANGLADESH SUBSET)

DATE	THEMES	LOCATIONS	PERSONS	ORGANIZATIONS	TONE	EVENTS	URLS
		Each record is data from a single news article	karnataka janata paksha;ganga kalyana;janata dal;sudhakar lal				
			saleemul huq;mary robinson;sheikh hasina;atiq rahman				
			andrew biraj;sirajul islam				
			ruhul amin gazi;baker hossain;mahmuda begum;abdus shahid;pran gopal dutta;ahmed shafi ahmed;ellias khan;pran gopal;khaleda zia;mohammad abullah;amar desh	*Multiple names of PERSONS are delimited by a ";"*			
			sheikh hasina;rana plaza				
			anbarasan ethirajan;rana plaza;mohammed asaduzzaman				

The modern study of network centrality as seen in human connections began under A. Bavelas at Group Networks Laboratory (Massachusetts Institute of Technology) in the late 1940's. Bavelas explored the connection between influence and structural centrality in small groups [8]. Today, we have hundreds of different techniques for measuring centrality in human connections, and these techniques have been applied to many different problem sets. While hundreds of network centrality measure exist, we explored our data using the four dominant centrality measures to evaluate regional influencer networks. We used *degree*, *closeness*, *betweenness*, and *eigenvector* centrality to establish which nodes in a network are the most important.

The most common network centrality measure is degree, which counts the number of neighbors, or edges, for each node in a network [9]. A node with a very high measured degree has many connections with other nodes. Degree centrality measures a node's involvement and activity within a group while ignoring the direction of the edges [10]. One of the primary benefits of degree centrality is its ease of computation given a network adjacency matrix. The negative side of using this measure is that connections between nodes do not always mean that those nodes are powerful or influential within their network. This is due to the fact that degree is a strictly local measure.

Betweeness, another centrality measure, identifies the nodes with the greatest number of "shortest paths" passing through it. The betweenness algorithm computes the shortest path between every node, and sums the number of paths that cross each node. The nodes with the highest number of shortest paths going through it are determined to be "central" to the network. Freeman defines betweenness mathematically in terms of the probability that point p_k falls on the random path between point p_i and point p_j [11]. If

$g_{ij}(p_k) =$ number of paths between p_i and p_j that contain p_k

then the probability that a random path contains p_k is

$$b_ij(p_k) = \frac{g_{ij}(p_k)}{g_{ij}}$$

The overall betweeness measure for point p_k is therefore the sum these probabilities as expressed by

$$C_B(p_k) = \sum_{j<k} \frac{g_{ij}(p_k)}{g_{ij}}$$

In our network analysis, this measure can be used to identify which people, places, or organizations create "bridges" in a network. Betweenness centrality is important for network analysis because it shows when one particular person or node is located in a strategically central position of the group. Such people in central positions can influence the group by "withholding information [or] coloring or distorting it in transmission" [11]. Since edge weights negatively affect the closeness, we removed weights for the closeness algorithm. This was done by simply removing duplicates in the edge list.

Closeness, another centrality measure, is the geodesic distance of average shortest path between a node and every other node in the network. Closeness can also be defined as the inverse of the farness from a node to all that a node is able to reach [12]. Closeness is a topological concept that says the closer two people are topologically, the greater influence they have on each other. To measure closeness of a point p_i, we define

$d\left(p_i, p_j\right) =$ the number of weighted edges between p_i and p_j

The closeness of point p_i can then be defined as

$$C_C(p_i) = \left[\sum_{j=i}^{N} d(p_i, p_j) \right]^{-1}$$

The closeness and betweenness algorithms both are computationally expensive since they rely on running Dijkstra's shortest path (or equivalent) algorithm $N(N-1)$ times. Like betweenness, edge weights negatively impact closeness and were removed for the closeness algorithm.

Eigenvector centrality, proposed by Bonacich in 1972, is a measure of the importance of a node in a network [13]. This method assigns relative scores to all the nodes in a network by giving "points" to nodes that are more influential than other nodes. Nodes in eigenvector centrality depend on both the numbers and the quality of their connections. While the number of connections contributes to a node's score, a node that has a small number of connections but a high quality will outrank other nodes with greater amounts of connections and mediocre quality [12]. Google's Page Rank is a variant of the eigenvector centrality measure. In our network analysis, connections to more influential people, as opposed to less influential people, will contribute to their own influence. Eigenvector centrality is mathematically defined as

$$\lambda \nu = A \nu$$

where A is the adjacency matrix, λ is the constant (the eigenvalue), and ν is the eigenvector.

While our final tool allows the user to use all four network centrality measures, we selected eigenvector centrality as the default. We found eigenvector centrality was computationally fast and fairly robust, even when sampling or boot strapping from large data sets. We also felt that the underlying premise of eigenvector centrality (i.e. weight ties by importance) closely matched what we felt commanders would value in their analysis.

C. The TABARI Algorithm

Our tool ingests data directly created on a daily basis by the TABARI algorithm. Text Analysis By Augmented Replacement Instructions (TABARI) is an algorithm that codes international event data from English news sources using pattern recognition and simple grammatical parsing. TABARI was created as part of the Penn State Event Data Project and is the open-source C++ successor of the NSF-funded Kansas Event Data System. The TABARI algorithm data has been used on several large projects, to include U.S. Department of Defense Intergrated Conflict Early Warning System (ICEWS) project [14].

There are two primary schools of thought among event data social scientists. The first is that we should have humans code events data by hand, erring toward small data sets with arguably increased accuracy (though several tests dispute the increased accuracy claim [14]). The other school of thought is that we should allow the computer to code large event data sets with some level of inaccuracy. While the purpose of this paper is not to defend computer generated event data sets, we chose to use a computer coded data set. Our reasons were more practical than theoretical; there does not exist a human-coded event data set with the volume and velocity to give daily updates across the globe. While we acknowledge that computer generated event data (as well as human generated event data), does contain errors, we believe that the nature of our project will allow strong and accurate signals (i.e influencers) to rise from the noise of the data.

III. Methodology

Our methodology focused on collecting and maintaining GKG data, and then querying, cleaning, analyzing, and presenting the results in a prototype Shiny web-application. This process is visualized in the Functional Flow diagram in Figure 2.

Due to the size and complexity of the Global Knowledge Graph we spent significant time managing, querying, and processing data. GKG is approximately 24 GB in daily comma separated value (CSV) files starting in April 2013. Data is added daily in the form of CSV files, and these daily files currently average 85,000 records. We used Unix programming on a cloud server to download daily updates every morning and store these in their original file structure. We then employed Unix programming to subset the data geographically, allowing us to create local, national, or regional networks. Most networks were build on the most recent 30 to 90 days of data. Using the R Programming language, we parsed the GKG subset, and built the adjacency matrix and/or edge list (we found the *edge list* computationally faster and used this data structure in our final solution). Persons that are found in the same article are connected to form an undirected graph. If two people are mentioned in a second article, the weight of their connection increases to 2 (we use a weighted adjacency matrix for degree and eigenvector, and un-weighted matrix for closeness and betweenness). Some articles, however, only contain one person. In order to still add value to these individuals, an edge is created connecting this person to themselves. The final weighted adjacency matrix is used for network centrality analysis.

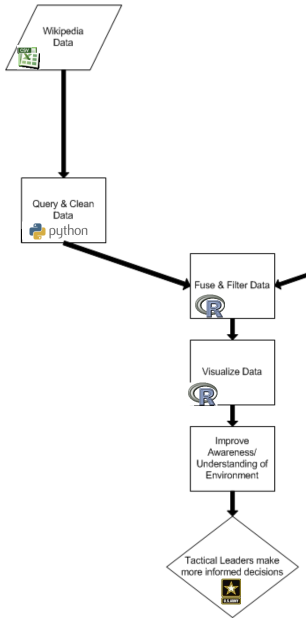
Wikipedia Data
Global Knowledge Graph Data
Query & Clean Data
Linux
Query & Clean Data
python
Analyze Network Centrality
Fuse & Filter Data
Visualize Data
Improve Awareness/ Understanding of Environment
Tactical Leaders make more informed decisions

Fig. 2. Functional Flow Diagram

We then apply all four network centrality algorithms (degree, betweenness, closeness, and eigenvector) to the network. Both *betweenness* and *closeness* were computationally expensive, since both rely on many iterations of a shortest-path algorithm. We decreased the computation time of betweenness by applying the *k-means betweenness* algorithm, and we decreased the computation time of closeness by applying the *k-means out closeness* algorithm. The centrality measures were merged and sorted to provide the user with the top n influencers in their selected region. This final list was rank ordered by the eigenvector centrality measure, our recommended default centrality measure. Once again, all network centrality calculations were conducted in the R Programming language. Finally, we employed a Python Wikipedia API in order to scrape a two sentence description for each individual. Depending on the region selected, Wikipedia contained descriptions for 30-100% of the top influencers in a region. This entire process was placed into a prototype web-application (produced with the RStudio Shiny package) and is currently being socialized with tactical commanders across conventional and special operations units.

Even using an edge list, there are times we encounter memory constraints due to the size of the networks (i.e. applying our tool to Washington D.C.). In these cases we conducted bootstrap sampling of the articles in order to derive influencer rankings. The default setting was 10 samples of 50,000 articles each. While conducting bootstrap sampling, we limited our network centrality measures to just degree and eigenvector for the sake of computational efficiency.

IV. Results

After developing and refining the model, we conducted network analysis of multiple city and national data subsets. Network summaries of many of these networks is given in Table II. Most networks contain from 1,000 to 180,000 nodes, and were very dense. As seen in the *Connectivity* field in Table II, every network is disconnected at some point. Note that Astana (Kazakhstan), Kinshasa (DRC), and Paraguay all have smaller networks. This is arguably due to the fact that they are not English speaking countries and were not in the news headlines in January 2015. That being said, however, even these three smaller networks provided valued information regarding influencers.

TABLE II. Network Statistics

Location	Vertices	Edges	Connectivity	Density	Diameter
dhaka	7019	65167	0	0.002646	11
estonia	3983	63339	0	0.007987	9
astana	1100	12820	0	0.021209	9
paraguay	1554	23515	0	0.019487	7
baghdad	9549	286691	0	0.006289	9
yemen	13741	533777	0	0.005654	8
liberia	10733	122168	0	0.002121	14
kinshasa	886	11876	0	0.030292	8
kabul	4088	65812	0	0.007878	9
paris	56317	1360881	0	0.000858	11
myanmar	5548	44176	0	0.002871	9

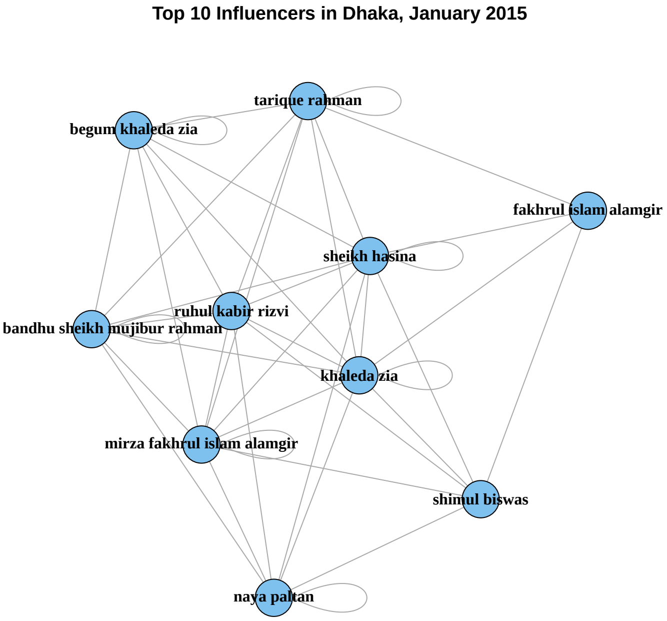

Fig. 3. Top 10 Influencers in Dhaka for January 2014 (7,074 total individuals in entire network)

A. Bangladesh Case Study

Using our tool we conducted a Case Study of Dhaka, Bangladesh. Dhaka was selected because it is of interest to the United States Pacific Command. Network centrality measures were compared based on computational time, accuracy, and value/insight to military commanders. The top 10 influencers for January 2015 are displayed in Table III. This chart is ordered by eigenvector rank, but provides the ranks of all other measures as well (note that our web-app allows the user to sort by a measure of their choice). This table was produced from a network of 7,074 individuals associated with Dhaka, Bangladesh. A plot of only the top 10 is provided in Figure 3 and the associated adjacency matrix is given in Table IV. Note the self-loop edges in Figure 3 indicating those individuals who were found in an article without any other persons mentioned. These individuals are highly connected at the center of the Dhaka network. Also note that the adjacency matrix is weighted by the number of co-mentions.

Throughout our observation the most influential people in Bangladesh were Sheikh Hasina and Khaldeh Zia, the current and former prime minister. We observed the interesting connections between the two and the similarities and difference in their connections. This information would alert a tactical leader to look more into the spheres of influence for each of these matriarchs. These two women have alternately controlled Bangladesh for the past 20 years. We were also able to see the way certain non-state actors worked within the social network. For example, in October 2014 we observed interesting connections between Ahmed Rajib, an atheist, anti-Islamic blogger, and Sheikh Hasina. After further research, we found that a group of extremist students murdered him in Dhaka in February 2013. Sheikh Hasina paid a visit to his family promising action in memory of their son. The story told is important information for tactical leaders. It helps to paint a picture of the social and cultural landscape in this region. Our results also helped us in locating key locations. The language algorithm in the GDELT interpretation did not recognize that Shahbagh Square is not a person, but a place. This mistake proved useful. Shahbag Square is very recent popular place for protests. It resides in close proximity to Dhaka University. Upon further research, we found that during 2013 Shahbag Square held many protests. One of these protests, which happened the night after Ahmed Rajib's murder, explains why the node shared connections with both Sheikh Hasina and Ahmed Rajib. [15].

TABLE IV. Weighted Adjacency Matrix for Top 10 Influencers in Dhaka

khaleda zia	144	603	180	143	71	29	87	68	71	67
sheikh hasina	603	52	107	55	81	102	81	39	57	68
tarique rahman	180	107	2	25	60	8	-	13	-	16
f. islam alamgir	143	55	25	14	8	2	19	35	34	-
mujibur rahman	71	81	60	8	4	4	-	7	1	-
begum khaleda zia	29	102	8	2	4	12	-	10	-	-
shimul biswas	87	81		19	-	-	-	21	20	28
ruhul kabir rizvi	68	39	13	35	7	10	21	0	27	0
naya paltan	71	57	-	34	1	-	20	27	2	-
islam alamgir	67	68	16	-	-	-	28	-	-	-

B. Baghdad Case Study

We additionally conducted a study of Baghdad due to the ISIS crisis in the region. During this case tested our method of randomly sampling data from a given time frame, building a network, determining primary influencers, and comparing this to a list of influencers developed from the entire network. In this case study, we randomly sampled 5,000 articles from a list of 9,000 articles. We then built influencer rankings for the sample and the entire population, and compared rankings using Kendall's Tau and Spearman's Rho rank correlation.

TABLE III. RESULTS FOR DHAKA, BANGLADESH (TOP 10 INFLUENCERS FOR JANUARY 2015)

’Name	Closeness Rank	Degree Rank	Betweenness Rank	Eigenvector Rank	Trending	Description
’khaleda zia	1	1	1	1	0	Begum Khaleda Zia (born 15 August 1945) is a Bangladeshi politician who was the Prime Minister of Bangladesh from 1991 to 1996 and again from 2001 to 2006. When she took office in 1991, she was the first woman in the country’s history and second in the Muslim world (after Benazir Bhutto of Pakistan in 1988-1990) to head a democratic government as prime minister. Khaleda Zia was the First Lady of Bangladesh during the presidency of her husband Ziaur Rahman.
’sheikh hasina	2	2	2	2	0	Sheikh Hasina (born 28 September 1947) is the current Prime Minister of Bangladesh, in office since January 2009. She previously served as Prime Minister from 1996 to 2001, and she has led the Bangladesh Awami League since 1981. She is the eldest of five children of Sheikh Mujibur Rahman, the founding father and first President of Bangladesh, and widow of the nuclear scientist M. A. Wazed Miah.
’begum khaleda zia	11	15	18	3	+3	Begum Khaleda Zia (born 15 August 1945) is a Bangladeshi politician who was the Prime Minister of Bangladesh from 1991 to 1996 and again from 2001 to 2006. When she took office in 1991, she was the first woman in the country’s history and second in the Muslim world (after Benazir Bhutto of Pakistan in 1988-1990) to head a democratic government as prime minister. Khaleda Zia was the First Lady of Bangladesh during the presidency of her husband Ziaur Rahman.
’arafat rahman koko	33	20	42	4	——-	Arafat Rahman, nicknamed ”Koko” alternative spelling ”Coco” (12 August 1970 - 24 January 2015) was the younger son of Former President of Bangladesh Ziaur Rahman and former Bangladeshi Prime Minister Khaleda Zia. Arafat was convicted in a money-laundering case by the Caretaker government of Bangladesh in 2007. On 17 July 2008, after taking permission from Bangladesh Supreme Court, he went to Thailand and from there to Malaysia for medical treatment.
’amir hossain amu	31	49	46	5	——-	Amir Hossain Amu is a Bangladeshi politician and senior leader in the Bangladesh Awami League.
’ziaur rahman	16	10	12	6	+8	Ziaur Rahman (19 January 1936 - 30 May 1981) was a Bangladeshi politician and first military dictator who served as the 7th President of Bangladesh from 21 April 1977 until his death in 30 May 1981. He was the announcer of the Bangladesh Liberation War on behalf of Bangabandhu.
’rafiqul islam	30	13	17	7	+9	Either Wikipedia Data is Not available for this Name or Multiple Wikipedia Answers are Available
’tarique rahman	22	9	24	8	-5	Tariq Rahman (also spelled as Tareq Rahman) (born: 20 November 1967) is a Bangladeshi politician. He is the current Senior Vice Chairman of Bangladesh Nationalist Party.
’saiful islam	34	33	27	9	——-	Mohammad Saiful Islam Khan (born April 14, 1969, in Dhaka) is a former Bangladeshi cricketer who played in 7 ODIs from 1990 to 1997. Originally from Mymensingh, Saiful played for the Bangladesh U-19 side in 1989. He made his full ODI debut at Eden Garden Calcutta on 31 December 1990. His best performance in ODI came at Sharjah against Sri Lanka in 1995. he took 4/36 to help Bangladesh bowl out their opposition for the first time in a full ODI. He also played in the Bangladesh side that won the ICC Trophy in 1997. But with the emergence of a group of talented medium pacers he soon lost his place in the national side.
’ruhul kabir rizvi	13	21	22	10	-3	Either Wikipedia Data is Not available for this Name or Multiple Wikipedia Answers are Available

TABLE V. RANK CORRELATION BETWEEN BAGHDAD SAMPLE AND POPULATION

	Degree	Betweenness	Closeness	Eigenvector
Kendall	0.87	0.88	1.00	0.96
Spearman	0.97	0.97	1.00	0.99

Our results are given in Table V. Note that the closeness and eigenvector centrality measures provide the most robust ranking measures in this case. In order to improve this even more, our tool bootstraps 10 samples when it encounters memory problems, therefore improving the accuracy of the influencer rankings.

V. CONCLUSION

After conducting both case studies, we concluded that this methodology and tool will provide value to military commanders and quickly shed relevant and timely light on a selected city, province, nation, or region. Network centrality measures were able to ascertain the signal among all the noise of machine generated event data. The fusion of Wikipedia to the GKG data allows a staff to quickly scan a list for individuals of interest, and research them further through OSINT or engagement.

We feel it is necessary to clearly articulate to military leaders two words of caution. First, although represented in a network topology, the networks that we generate are not social networks. While some social connections exist in our networks, the fact that two individuals are in the same news article does not in any way define a social connection. Our methodology is meant to rank order influencers in a region, not define a social network topology. Second, the GKG data set is not a truly random sample of news and events around the world, and is therefore not evenly distributed. The first and obvious bias of GKG data is that it currently only codes English new sources, and is therefore biased toward English speaking cultures and societies. Additionally, news agencies, by their very nature, do not present a random sample of events in any region, but are generally biased toward population, cultural, and political centers of gravity. Events and therefore influencers in rural, non-English speaking areas are not apt to appear in news articles, and even if they do they may disappear amongst the noise of GKG.

VI. FUTURE WORK

Given the success of rank ordering influential persons in a region, we would like to apply the same methodology to organizations (found in a different field in GKG data). We feel that providing a military leaders with a list of both influential individuals as well as influential organizations would provide tremendous value. We are in ongoing discussions with military leaders regarding whether or not to model individuals and organizations in the same network or separate networks.

We are currently socializing our tool with conventional as well as special operations forces in the US Army. The tool has also gained significant interest among Civil Affairs units as well as the U.S. Army War College at Carlisle Barracks. Additionally, we are collaborating with ongoing efforts at Yale University as well as the United States Institute of Peace.

Acknowledgment

The authors would like to thank the Data Tactics Corporation for their support and collaboration throughout this project.

References

[1] D. Lamothe, Ed., *U.S. Military force fighting Ebola virus could grow up to 4,000 troops*, Washington Post, October 03 2014.

[2] "Regionally aligned forces and global engagement," in *Contemporary MIlitary Forum III*. U.S. Army, October 2013, AUSA Conference.

[3] K. C. Walker, *U.S. Army Functional Concept for Engagement*, U.S. Army Training and Doctrine Command, February 2014.

[4] *ADRP 2-0: Intelligence*, Headquarters, Department of the Army, August 2012.

[5] K. Leetaru and P. A. Schrodt, "GDELT: Global data on events, location, and tone, 1979–2012," in *ISA Annual Convention*, vol. 2, 2013, p. 4.

[6] U. Army, "Army doctrine reference publication (ADRP) 5-0," *The Operations Process*.

[7] K. Börner, S. Sanyal, and A. Vespignani, "Network science," *Annual review of information science and technology*, vol. 41, no. 1, pp. 537–607, 2007.

[8] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social networks*, vol. 1, no. 3, pp. 215–239, 1979.

[9] R. Diestel, *Graph Theory*. Graduate Texts in Mathematics, 2012.

[10] C. Prell, *Social Network Analysis: History, Theory and Methodology*. SAGE Publications, 2012.

[11] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry, Vol. 40, No. 1 (Mar., 1977), pp. 35-41*, 1977.

[12] M. Newman, "A measure of betweenness centrality based on random walks," *Department of Physics and Center for the Study of Complex Systems, University of Michigan, Ann Arbor, MI 48109*, 2003.

[13] P. Bonacich, "Some unique properties of eigenvector centrality," *Social Networks*, vol. 29, no. 4, pp. 555–564, 2007.

[14] P. A. Schrodt and J. Yonamine, "Automated coding of very large scale political event data," in *Presentation to the New Directions in Text as Data Workshop, Harvard University, October*, 2012.

[15] E. Barry, Ed., *Matriarchs' Duel for Power Threatens to Tilt Bangladesh Off Balance*, New York Times, January 11 2014.

[16] *FM 3-22: Army Support to Security Cooperation*, Headquarters, Department of the Army, January 2013.

[17] *ADRP 3-0: Unified Land Operations*, Headquarters, Department of the Army, May 2012.