Military Cyber Affairs
The Journal of the Military Cyber Professionals Association
ISSN: 2378-0789

Volume 2 | Issue 1 Article 3

2016

Leveraging Lexical Link Analysis (LLA) To Discover New Knowledge

Ying Zhao
Naval Postgraduate School, yzhao@nps.edu

Douglas J. MacKinnon
Naval Postgraduate School, djmackin@nps.edu

Shelley P. Gallup
Naval Postgraduate School, spgallup@nps.edu

Joseph L. Billingsley
Naval Postgraduate School, jbilling@nps.edu

Recommended Citation

Zhao, Ying; MacKinnon, Douglas J.; Gallup, Shelley P.; and Billingsley, Joseph L. (2016) "Leveraging Lexical Link Analysis (LLA) To Discover New Knowledge," *Military Cyber Affairs*: Vol. 2 : Iss. 1 , Article 3.
DOI: http://doi.org/10.5038/2378-0789.2.1.1015
Available at: http://scholarcommons.usf.edu/mca/vol2/iss1/3

Leveraging Lexical Link Analysis (LLA) To Discover New Knowledge

Cover Page Footnote
We would like to thank Professor David Trask at the Naval Postgraduate School for his guidance. We also like to thank Dr. Charles Zhou from Quantum Intelligence, Inc. for his valuable input and discussion related to Collaborative Learning Agents (CLA).

Leveraging Lexical Link Analysis (LLA) To Discover New Knowledge

YING ZHAO, Naval Postgraduate School
DOUGLAS J. MACKINNON, Naval Postgraduate School
SHELLEY P. GALLUP, Naval Postgraduate School
JOSEPH L. BILLINGSLEY, Naval Postgraduate School

Lexical Link Analysis (LLA) is a form of text mining in which word meanings represented in lexical terms (e.g., word pairs) are treated as if they are in a community of a word network. LLA can provide automated awareness for analyzing text data and reveal previously unknown, data-driven themed connections. We applied LLA to develop authentic archetypes and conversely determine potential imposters of that same archetype. We use publically available social media data to develop a cyber professional as an example. This paper reports the development of the algorithm, the collection and labeling of data, as well as the results of analysis of the characteristics of three authentic cyber professionals using data collected from Facebook, LinkedIn and Twitter. This method can provide automated analyzing and understanding massive Big Data in open and social media data sources to discover new knowledge for a widening range of applications.

1. Introduction

The objective of this research is to discover features that are potential indicators for evaluating authenticity of an entity, *e.g.*, a person or an organization, through automated collection and analysis of public social media profile data.

Authenticity is a technical term used in psychology as the degree to which one is true to one's own personality, spirit, or character, despite external pressures. Authority is someone having or showing impressive knowledge about a subject or someone who is respected or obeyed by other people.

Why is authenticity important? Why do we do the research? Public platforms (*e.g.*, publications) and social media have provided opportunities for individuals and organizations to establish their personas that are perceived as authentic and authoritative, however, are they truly authentic and authoritative? As an example, we began by asking how the authentic persona of *a rocket scientist from the Naval Postgraduate School (NPS) in Monterey, California* might appear in social media. As we began to look through Facebook, such a person might choose to associate with particular professional organizations such as NASA or NPS. We also asked how a resident of the Monterey Peninsula might appear. An authentic person may, for instance, include personal interests and likes such as 17-Mile Drive, which is located near NPS. The same person may also have many followers and likes for posts related to local events.

Three research questions emerged:

1. How can we discover and learn the characteristics of authentic personas across multiple data sources and platforms?
2. How can we apply these discovered characteristics to evaluate a new claimed persona and assess its authenticity?
3. Can this process be automated?

Traditionally in social networks, the importance, authority, or authenticity of any network node, for example, a leadership role in a social network [1] [2] [3] is measured according to various

[1] Blei, D., Ng, A. and Jordan, M. 2003. Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993-1022. Retrieved from http://jmlr.csail.mit.edu/papers/volume3/blei03a/blei03a.pdf

centrality measures[4]. Among a large collection of centrality measures, sorting and ranking information (*e.g.,* keywords or features for web pages or blogs in the internet[5]) of authority or authenticity is compared with page ranking of a typical search engine. Current automated methods such as graph-based ranking used in PageRank[6], require established hyperlinks, citation networks, social networks (*e.g.,* Facebook), or other forms of crowd-sourced collective intelligence. Similar to the PageRank algorithm, HITS algorithm[7], TextRank[8] and LexRank[9] have been used for keyword extraction and document summarization. These methods first build a graph based on the similarity of relationships among nodes which can represent people, words (for keywords extraction), or sentences (for document summarization). The importance, authority, or authenticity of each node is determined by computing an authority score that equals the number of nodes pointing to the node.

However, these methods are not applicable to situations where there are no pre-established relationships among network nodes. For example, there are no hyperlinks available in non-structured or public social media data and many DoD Big Data. This makes the traditional centrality measures or PageRank-like methods difficult to apply. Furthermore, current methods mainly score popular information and do not rank emerging and anomalous information which might be more important in many applications.

In this paper, we seek to show that our discovered correlations may reveal matches and deviations, leading to emerging and anomalous information, and resulting in varying degrees of authenticity. We collected and analyzed public social media data to define personas and profiles, and to discern between authentic and false archetypes using social media data to develop patterns and trends learned and trained from authentic personas to validate our findings. We also show how Lexical Link Analysis (LLA) may be used to discover interesting themes and establish authoritative and authentic personas, or profiles. We have shown previously [10] [11] [12] [13] that LLA results are highly correlated with manual link analysis done by human analysts.

[2] Center for Computational Analysis of Social and Organizational Systems (CASOS) 2009. AutoMap: Extract, analyze and represent relational data from texts. Retrieved from http://www.casos.cs.cmu.edu

[3] Girvan, M., and Newman, M. E. J. 2002. Community structure in social and biological networks. Proceedings of the National Academy of Sciences, USA, 99(12), 7821–7826

[4] Freeman, L.C. 1979. Centrality in social networks I: Conceptual clarification. Social Networks, 1: 215-239

[5] Marlow, C. 2004 Audience, structure and authority in the weblog Community. 54th Annual Conference of the International Communications Association International Communication Association Conference, May, New Orleans, LA

[6] Brin, S. and Page, L. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems, 30:107-117

[7] Kleinberg, J. M. 1999. Authoritative sources in a hyperlinked environment. Journal of the ACM 46 (5): 604. http://en.wikipedia.org/wiki/HITS_algorithm

[8] Mihakcea, R. and Tarau, P. 2004. TextRank: Bringing order into texts. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), July 2004, Barcelona, Spain

[9] Erkan, G. and Radev, D. R. 2004. Lexrank: graph-based lexical centrality as salience in text summarization. J. Artif. Int. Res., 22:457–479

[10] Zhao, Y., Gallup, S. P., and MacKinnon, D. J. 2010. Towards real-time program-awareness via lexical link analysis. Proceedings of the Seventh Annual Acquisition Research Symposium. Retrieved from http://acquisitionresearch.net

[11] Zhao, Y., Gallup, S.P. and MacKinnon, D.J. 2011a. System self-awareness and related methods for improving the use and understanding of data within DoD. Software Quality Professional, 13(4): 19-31. http://asq.org/pub/sqp/

[12] Zhao, Y., Gallup, S.P. and MacKinnon, D. J. 2011b. A web service implementation for large-sale automation, visualization and real-time program-awareness via Lexical Link Analysis. Proceedings of the Eighth Annual Acquisition Research Program, Monterey, CA: Naval Postgraduate School

2. Methods

2.1 LLA Method

We frame our thinking about words connected in a network as in military operations, where the term *situational awareness* is coined to help improve a decision maker's understanding of their surrounding environment. Using LLA, we extend the concept of awareness as the cognitive interface between decision makers and a complex system. A complex system can be expressed in a list of *attributes* or *features* which are specific vocabulary or lexicon terms to describe its characteristics and surrounding environment.

Specifically, LLA is a form of text mining. For example, word pairs or bi-grams as lexical terms and features can be extracted and learned from a document repository. LLA automatically discovers word pairs or bi-grams and displays them as a network from data. In a text data set, words form the nodes and pairs of words or bi-gram word pairs form the links between nodes. Fig. 1 shows an example of such a word network discovered using LLA; "middle_east analyst", "intelligence analyst", and "energy analyst" are examples of bi-gram word pairs discovered from data.

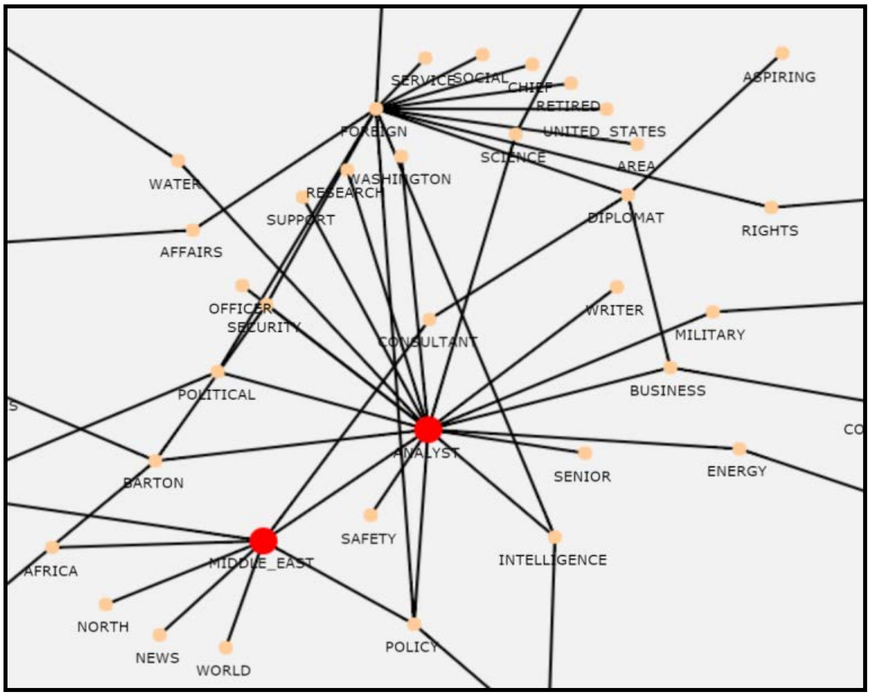

Fig. 1. Example of bi-gram based word pair networks.

[13] Zhao, Y., Mackinnon, D. J., Gallup, S. P. 2015. Big data and deep learning for understanding DoD data. Journal of Defense Software Engineering, Special Issue: Data Mining and Metrics

The details of LLA processing include the following steps:

Step 1: Filter out a list of pre-defined stop words; for example, the words “a”, “the”, “this” and “that,” which do not convey meaning in English. Select word pairs in a sentence or paragraph level based on the bi-gram parameters of the following:

- The probability of one word next to another word
- The minimum frequency for each individual word

Step 2: Apply a social network community finding algorithm, (*e.g.*, the Newman community detection method, to group the word pairs into *themes or topics*.) A *theme* includes a cluster or community of word pairs connected to each other.

Step 3: Compute an *importance* measure for each theme.

Step 4: Sort *theme importance* measured by time or other sequential parameters, and study the distributions of the discovered *themes*.

This method is shown in the following example and visualization. Fig. 2 depicts LLA findings using connected groups or *themes*. Words are linked as word pairs that appear next to each other in the original documents. Different colors indicate different clusters. Each word pair cluster is produced using a social network community detection method, where words are connected and grouped, as shown in a single color, as if they are in a social community. A word center is formed around a word node connected with a list of other words in word pairs.

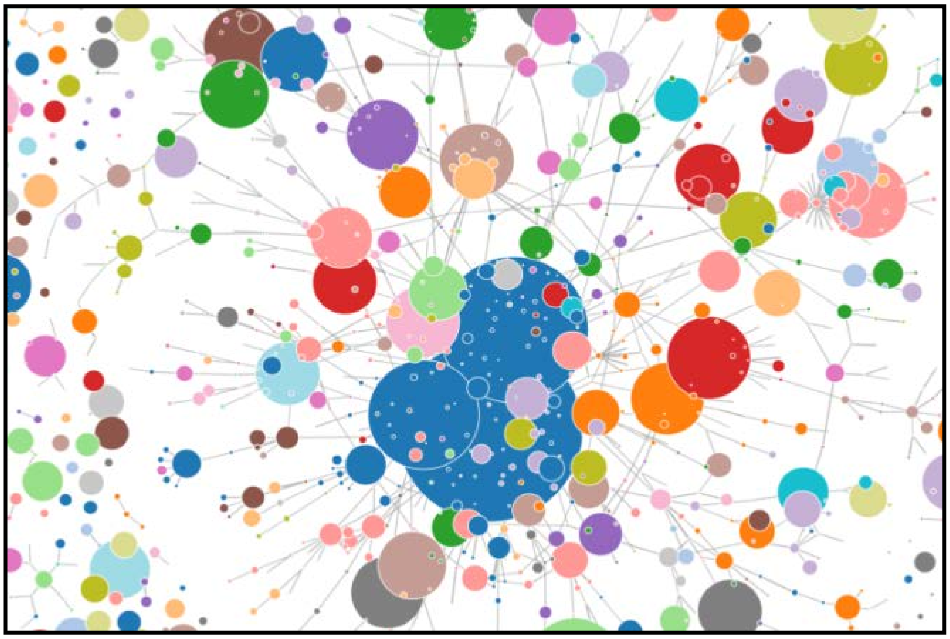

Fig. 2. Themes discovered and shown in colored groups. A node represents a word in a corpus and a link or edge represents a word pair. Each color of the links represents a theme.

Fig. 3 shows a detailed view of a *theme* in Fig. 2. The center words are “university, college, high_school.” In this example, we use a three-word *theme* of “university, college, high_school” to label such a group: the top three words with the highest *total degree of centrality*. More than three central word nodes can be chosen for labeling a group. The meaning of a theme is usually verified and validated by a human analyst.

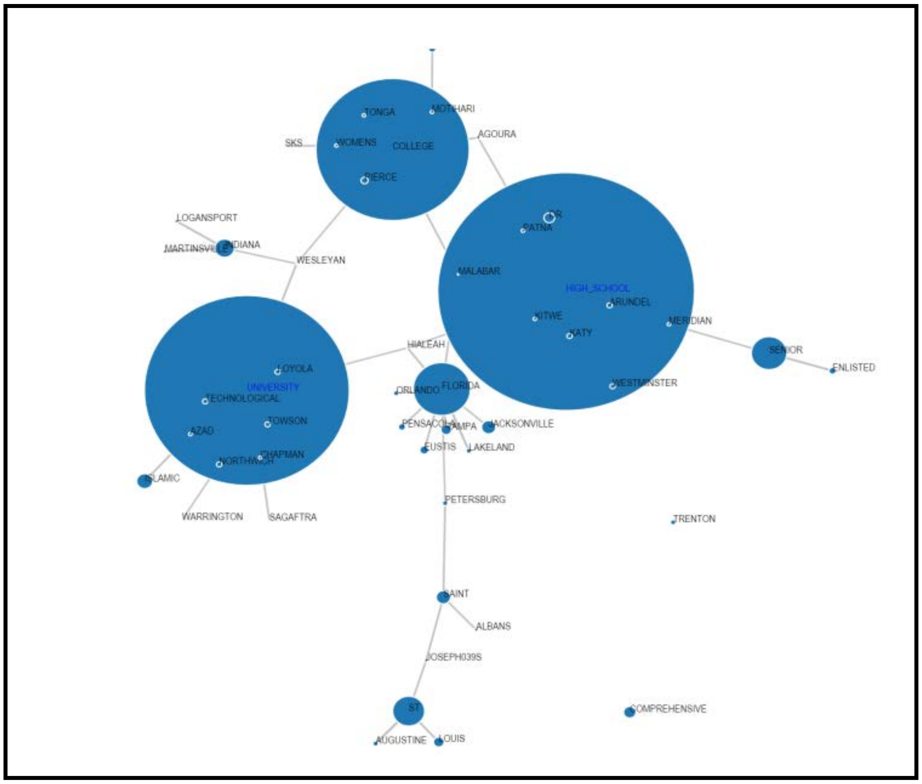

Fig. 3. A detailed view of a theme in Fig. 2

Equipped with word pairs or bi-grams as basic meaning units, and themes as the next level meaning units, LLA is often used to compare two data sets or two document sets where some word pairs or themes are shared. A system, or a corpus, can be a collection of documents for an actual physical system.

Fig. 4 shows a visualization of common *lexical links* that are shared between Systems 1 and 2, captured within the added red box. These connected links between systems reveal what these two systems share in common. Unlinked, outer vectors (outside the red box) indicate unique system features found individually for each system.

The closeness of the systems in comparison can be examined visually or using the quadratic assignment procedure (QAP[14]) to compute the correlation of two sets of lexical terms from two systems and to analyze the structural differences in the two systems.

[14] Hubert, L. and Schultz, J. 1976. Quadratic assignment as a general data-analysis strategy. British Journal of Mathematical and Statistical Psychology, 29: 190-241

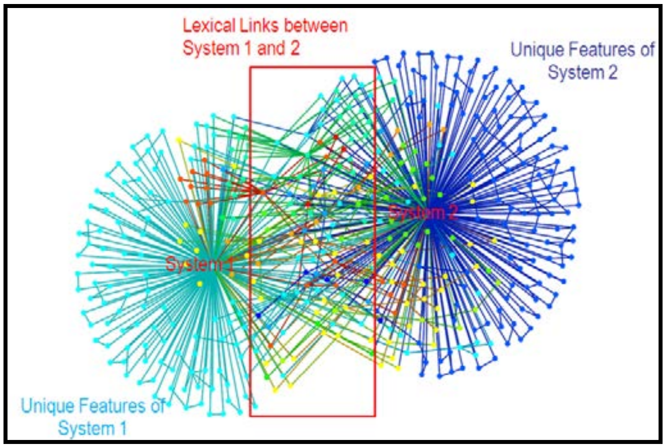

Fig. 4. Comparing two systems using LLA

LLA is related to the bags-of-words (BAG) method such as Latent Semantic Analysis (LSA)[15], Probabilistic Latent Semantic Analysis (PLSA)[16], WordNet[17], Automap[18], and Latent *Dirichlet* Allocation (LDA)[19]. LDA uses a bag of single words (*e.g.*, associations are computed at the word level) to extract concepts and topics. Alternatively, the idea of using the graph, or Text-As-Network (TAN), to model text, for example, allows extracting keywords used in text summarization. The TAN methods are not new, yet the TAN approach does, however, provide analysis over and above LDA and other BAG methods, since network theories – such as community detection methods, node ranking methods, and network growth theories – can be readily used to cluster and prioritize features. Since our approach learns the syntax, grammars and synonyms from data (*e.g.*, the word pairs that appear in the same context together), and discovers word pair clusters in a latent space, it is not sensitive to the choice of words that could be different among different people. The semantic spaces and latent information are represented using the word pair clusters or themes, since each cluster includes word pairs that are correlated to each other, even synonyms. Each word pair cluster was produced using a social network community detection method. LDA might appear more robust to such choices when it uses a bag of words instead of word pairs; however, word pairs in our approach represent more meaningful features than a bag of words (see more discussion in Section 4). Among the TAN methods, there are two types of term relationships: *term co-occurrence* and *term dependency*. In a *term dependency* network, the networks are constructed from pre-defined syntactic dependence, *e.g.*, one word is always subordinated (dependent) to the other syntactically[20]. Stanford Lexical Parser (SLP) [21] is a dependency-parsed text network.

[15] Dumais, S. T., Furnas, G. W., Landauer, T. K. and Deerwester, S. 1988. Using latent semantic analysis to improve information retrieval. In Proceedings of CHI'88: Conference on Human Factors in Computing, 281-285

[16] Hofmann, T. 1999. Probabilistic Latent Semantic Analysis. Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, Stockholm, Sweden

[17] Miller, G. A. 1995. WordNet: a lexical database for English. Communications of the ACM, 38(11)

[18] Center for Computational Analysis of Social and Organizational Systems (CASOS) 2009. AutoMap: Extract, analyze and represent relational data from texts. Retrieved from http://www.casos.cs.cmu.edu

[19] Reference 1

[20] Otero, P. G. 2012. The Meaning of Syntactic Dependencies. Retrieved from http://www.linguistik-online.de/35_08/gamallo.html

[21] The Stanford Natural Language Processing Group (SNLP). 2012. The Stanford Parser: A statistical parser. Available: http://nlp.stanford.edu/software/lex-parser.shtml

DOI: http://doi.org/10.5038/2378-0789.2.1.1015

In a term co-occurrence network, two terms or nodes are connected if they occur within the same textual context, (*e.g.,* next to each other (bi-gram), within *n*-word windows (*n*-gram), in a same sentence, same paragraph, and so on). LLA is a TAN method that uses word pairs, bi-gram networks, and network communities, to model topics and concepts (themes). LLA uses bi-gram word pairs, which can be extended to *n*-gram terms, as the basis to form word networks instead of just a statistically analyzed *bag of words*. *N*-gram, as a pre-processing step, reduces the total number of features or nodes in a word network, in addition to other pre-processing steps such as pre-identifying parts of speech (POS), (*i.e.*, tagging words in terms of noun, verb, adjective[22]) or stemming (*e.g.,* reducing derived words to their word stem). *N*-gram analysis allows LLA to focus only on the *important* terms that pass the *n*-gram selection criteria. For example, with a pre-specified probability threshold *t*, LLA automatically discovers *n*-grams (*n*>2) that match the threshold (*e.g.,* phrases with two or more words that have the probability larger than the threshold, for example, in Fig. 1 "middle east" is a phrase matched a threshold, therefore is connected using a "_").

2.2 Learning Authentic Archetypes

By leveraging text analysis, and in conjunction of LLA, we can consider how data are categorized, ranked, and sorted in globally interplayed social and semantic networks. After initial data clustering where we identify word pairs (bi-grams), we next consider its importance among the bi-grams and use the bi-grams for a complex system as *features*. We divide the *features* into three sections: authoritative or patterned features, emerging features, or anomalous features.

Each cluster, although discovered using a modularity measure, is further validated using statistical hypothesis testing, where a *p*-value is generated and compared to ensure statistical significance, specifically:

- Authoritative or Popular (P) word pairs: clusters or themes containing the highest number of mutually connected word pairs or features. These represent the main topics in a corpus at present. The data represented in Fig. 3 is an example of a popular theme centered on the word nodes "university, college, high_school". They can be insightful in two ways:
  - Authoritative: These terms may be shared or cross-validated across multiple diversified sources, so they are considered "authentic" or "authoritative".
  - Popular: These themes could be less interesting because they are already in the public consensus and awareness.
- Emerging (E) word pairs: Clusters or themes containing the intermediate number of mutually connected word pairs – these themes may grow to become popular or authoritative over time as our analysis continues. The theme represented in Fig. 5 is an example of an emerging theme centered on the word nodes "management, VA, Virginia".
- Anomalous (A) word pairs: Clusters or themes containing the lowest number of mutually connected word pairs. These themes may not seem to belong to the formed cluster as compared to others yet may be interesting for further investigation. Fig. 6 is an example of an anomalous theme for data centered on the word nodes "science, police, officer."

[22] Toutanova, K. and Manning, C. 2000. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), Hong Kong

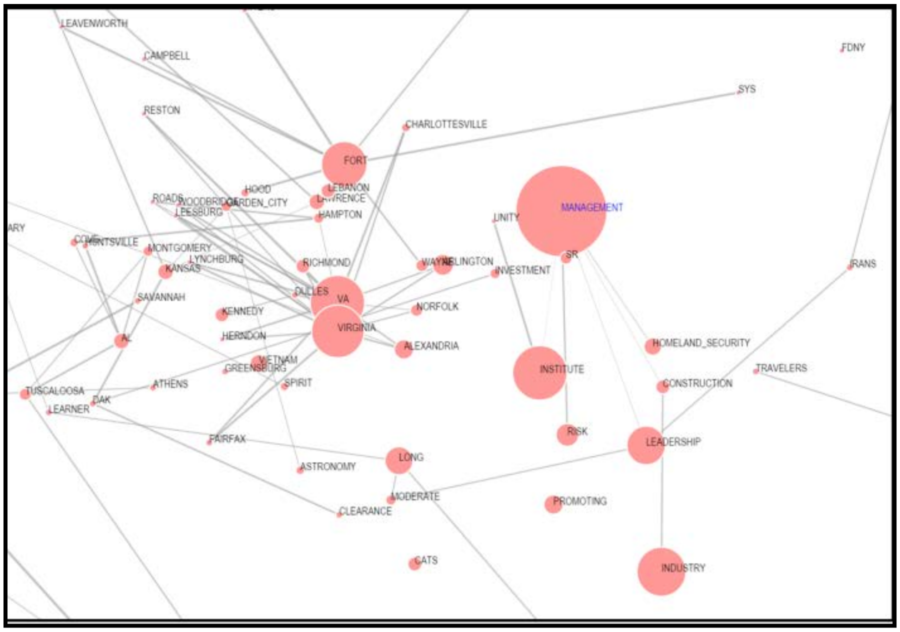

Fig. 5. An example of emerging theme

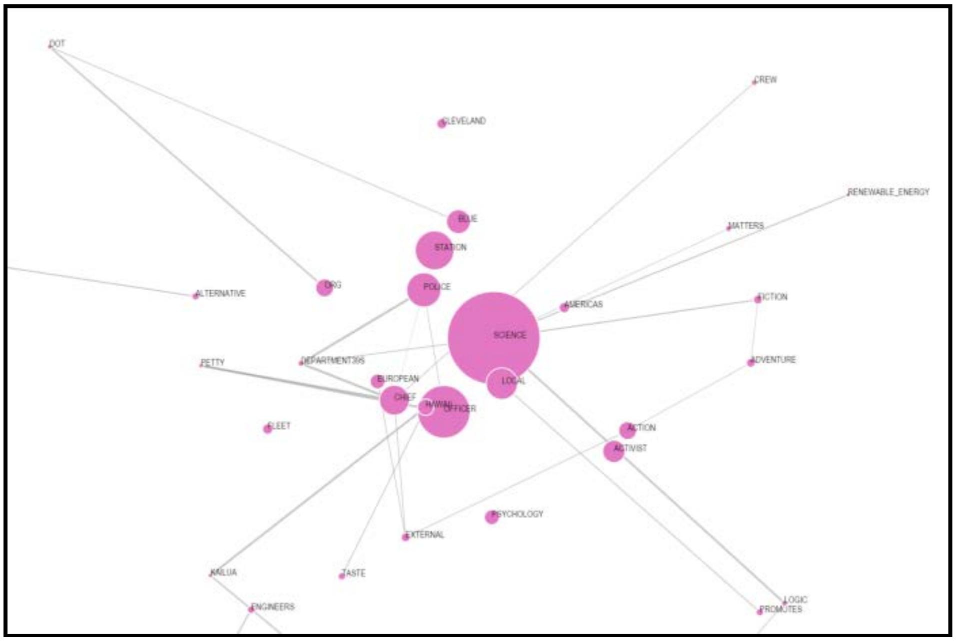

Fig. 6. An example of anomalous theme

With regard to our research questions, the methodology for determining authenticity using LLA is illustrated in Fig. 7. In Fig. 7, once authentic persona 1 is defined, the difference between a new persona 2 and the authentic persona 1 can be indicators for an anomalous persona. Our example shown is derived from developing an authentic archetype for a *cyber* professional gained from multiple samples of social media data. Authentic archetypes are developed from the feature clusters of authoritative, emerging, and anomalous clusters discussed above. This archetype development leads to a quantitative correlation that can be computed to determine relative authenticity for a new person with a similar public profile.

The LLA steps for the specific case of developing "cyber professional" archetypes are detailed as follows:

Step 1: Collect data. We began by identifying a list of public Facebook, Twitter, and LinkedIn pages – which appeared to be interesting to cyber professionals – and collected public profiles of people who made comments on these sites. We only collected public profiles.

Step 2: Identify a few seed "authentic documents" from a domain expert. These documents provided initial authentic vocabularies and lexical terms in the cyber field that can be used to match the online public profiles collected from Step 1.

Step 3: Generate a match matrix using LLA from the public profiles collected from Step 1 to the authentic documents, revealing how well specific profiles are matched with authentic vocabularies in the seed documents.

Step 4: Determine the top *n* matched profiles, since top profiles are considered more likely to be authentic based on discovered LLA scores found through matched features. Iterate *Step 2* and add the top *n* matched profiles as authentic documents.

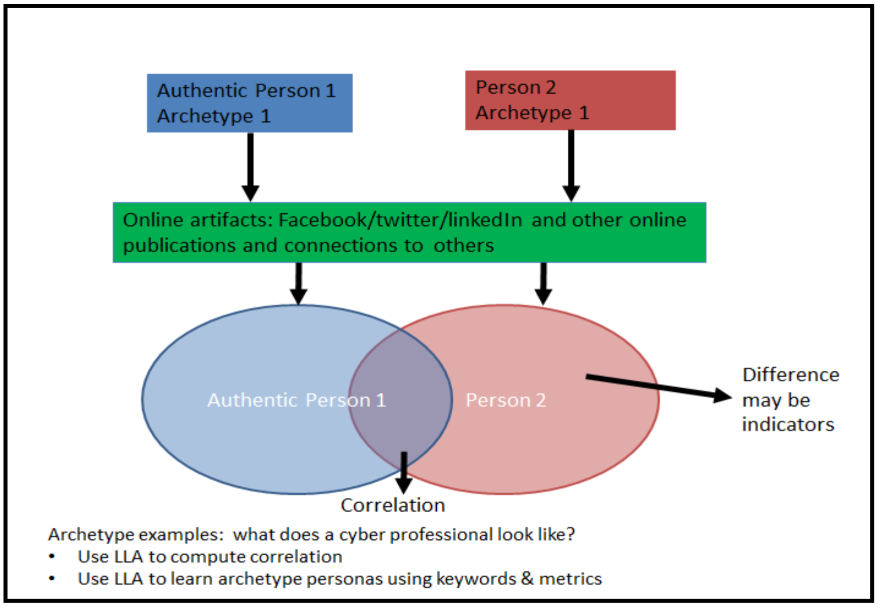
Authentic Person 1
Archetype 1
Person 2
Archetype 1
Online artifacts: Facebook/twitter/linkedIn and other online publications and connections to others
Authentic Person 1
Person 2
Difference may be indicators
Correlation
Archetype examples: what does a cyber professional look like?
• Use LLA to compute correlation
• Use LLA to learn archetype personas using keywords & metrics

Fig. 7. Correlation of online features for an authentic persona.

3. Results

Results (statistics and findings) reported below are based on publically available data to derive online characteristics or personas of authentic *cyber* professionals.

- We collected data from a total of 443 (394 Facebook, 48 Twitter and one Linkedin) websites for a period of six months.
- There were 10165 (1619 Facebook profiles, 8395 Twitter profiles, and 151 Linkedin profiles) collected.

From this:

- 1352 profiles were *matched* with one or more lexical terms (*e.g.*, word pairs) to a set of four authentic cyber professional documents used for an LLA model as detailed in Section 2.2
- 303 authentic profiles here, clustered into three archetypes as shown in Table 1 below.

- 38 anomalous profiles

In addition to the text of comments and original posts to which the comments were made, we also included other numerical or categorical features of profiles and discretized them into vocabularies, (*e.g.*, the number of comments made within the time period (6/2013 to 11/2013), the number of locations or screen name changes, and the numbers of emerging or anomalous features categorized by LLA). These features were used together with other keywords matched with the seed documents to discover the archetypes listed in Table 1. There were three archetypes discovered. The number of archetypes was not predefined.

Table 1.Statistics of *Matched* Profiles Collected from 6/2013 to 11/2013

	Number	**Characteristics**
Authentic Personas Archetype 1	245	• Speak using domain expert vocabularies • None or few location changes, • Consistent/frequent comments and Twitter names active
Authentic Personas Archetype 2	42	• One comment but associates with authentic organizations
Authentic Personas Archetype 3	16	• One comment, but uses emerging and domain expertise terms
Anomalous Profiles	38	• None or very few lexical links to the authentic documents • Frequent change of addresses/screen names • Blocked Twitter names
Inconclusive Profiles	1011	• One comment, not enough data to decide

3.1 Characteristics of Anomalous and Authentic Personas

We identified numerous themes and resulting consistent features that describe “cyber professional” archetypes as discussed in the next subsection below. We performed a face validation using known “cyber professional” profiles and found that our genuine “cyber professionals” were included in the clusters of our archetype features.

Conversely, when we compared these features to all those claiming to be “cyber professionals,” we also found individuals with attributes that appeared inconsistent with our archetype features. Those individuals with low matching scores exhibited multiple LinkedIn addresses, multiple screen names, multiple Twitter names, disparate, infrequent comments, and blocked or briefly existing Twitter names. Our known authentic personas, conversely, demonstrate the opposite behaviors. Behaviors that authentic personas demonstrated in the social media public profiles are listed as follows:

Authentic Personas Archetype 1:

“Speaks the lingo”: We found authentic cyber professionals have more authentic vocabularies in the expertise domain (*i.e.*, more matches to the “authentic documents” which were the four documents provided initial authentic vocabularies and lexical terms in the cyber field which were used to match with online public profiles).

- Had no (or very few) location, address, job description changes for a time period (*i.e.*, six months)
- Had consistent screen names and twitter names
- Had consistent, frequent, and germane comments
- Had persistent Twitter names which were not currently blocked

We also found two other archetypes of authentic personas:

Authentic Personas Archetype 2:

"Linked": These contained only one comment or one public activity during the time period (*i.e.*, not Archetype 1); however, these public profiles associated themselves (*e.g.*, referenced online profiles, emails, employment or memberships) with authentic organizations germane to their persona. Authentic organizations typically have large numbers of public "likes" or "followers" in Facebook, LinkedIn, and Twitter.

The data collection can also extend to other online sources. Different professional organizations and communities might give different levels of credibility in terms of online references, for example, a reference from NASA would be stronger than a LinkedIn connection which exhibits self-determined input. Publications in conferences and professional publication organizations are considered credible references.

Authentic Personas Archetype 3:

"Latency": These contained only one comment or one public activity between June and November (*i.e.*, not Archetype 1 or 2); however, these profiles used domain-expert terms related to, but not in, the authentic documents. The following list contains these types of related terms as captured by LLA:

- SEO (search engine optimization), SEM (search engine marketing)
- smart IT, information sharing
- air-to-air or SAM (surface-to-*air* missile) capabilities
- secure computing
- ipad/iphone/video game, transmedia special ops
- field operations
- sensor buoys/USV with sonars and powerful optics, detect subs
- full spectrum voice, VoIP, Internet, data, and cloud provider
- real-time situational awareness
- logistics support
- air conditioning integrity
- demand response, smart grid

3.2 Resolutions of Same Persons

When studying authenticity in social media data, we observed that the same person may have multiple different profiles. A research question emerged: What is needed to resolve different profiles into same persons using features or attributes discovered using LLA?

In an extended data set collected from 6/2013 to 1/2014, 307 out of 18651 (1.6%) profiles have the same "first and last" combinations, yet multiple profiles in different platforms such as Twitter, Facebook, Linkedin or multiple screen names in the same platform such as Twitter.

If two profiles, with same last name, same first and possible different middle name, however, are matched with one or more attributes or keywords (*e.g.*, location), LLA predicts two profiles are the “same person”; otherwise, “different person”. These features are in the “emerging” and “anomalous” categories.

Among the 307 profiles, we divided the profiles into the following cases:

- Case 1: Resolve to the same person due to identical profiles: 170
- Case 2: Resolve to the same person using *matched features* (*e.g.,* key words to describe interests & hobbies (*e.g.,* founder/entrepreneur, cyber security, afcea/signal), locations (WA, HI, Kansas) and affiliations(CIMSEC, George Washington University): 37
- Case 3: Resolve to the different persons due to lack of matched features: 74
- Case 4: Un-resolved or profiles deleted currently: 16
- Case 5: Organizations: 10

Therefore, the rate of resolving to the same person is (37+170)/307=67%.

We also defined a “popular last name” as a last name shown with more than one unique first names; and a “not popular last name” as a last name shown only once with other names (first or middle). Non-popular names tend to resolve to the same person.

4. Discussion

4.1 Possible Applications

By using LLA, words or features of people are represented in word pair networks. These networks can be analyzed, for example, using centrality measures as in real-life social networks, such as network theories, (*e.g.,* centrality measures, which typically measure the position and influence of a node in a social network), can be readily depicted to evaluate the importance of lexical terms in the word networks. The computation of *importance* of a node, or the authority score, takes into account the global information on a graph which can recursively be transformed into the problem of solving the principal eigenvector of a transition matrix. PageRank for the current internet search engine is based on this concept. The authority score also is related to the preferential attachment growth pattern[23] [24]; namely, some information is valuable or more important because it links to existing important information or systems. On the contrary, the expertise score is rarely studied and represents our innovation here. It measures how unique a system is that possesses some information that no other systems have.

In a previous research, we showed that the authenticity and authoritative method we studied here is highly correlated with typical research citations, or *h*-index[25], of authority measures when ground truth was given[26]. Therefore, the method can be used as an alternative approach for research citations or *h*-index because the method does not need citation data. Furthermore, this method can be used to improve and compute more accurate research citations or *h*-index when citation data are available. For example, when people use names differently, sometimes with and sometimes without middle names, with first names – and sometimes with first name initials – one can apply the method stated in this paper to link the same people based on public data via

[23] Barabási A.L. and Albert R. 1999. Science 286: 509

[24] Borgs, C., Chayes, J., Daskalakis, C. and Roch, S. 2007. First to market is not everything: an analysis of preferential attachment with fitness. Proceedings of the thirty-ninth annual ACM symposium on Theory of computing. ACM New York, NY, USA. Retrieved from http://research.microsoft.com/en-us/um/people/borgs/papers/fit.pdf

[25] https://en.wikipedia.org/wiki/H-index

[26] Reference 10

publications and social media data, and then apply traditional research citation or *h*-index methods.

4.2 Automation

With respect to Question 3 stated in the introduction, we implemented the data collection and the LLA method using a smart infrastructure called Collaborative Learning Agents (CLA)[27]. A learning agent is a domain expert which holds certain unique domain knowledge or expertise. At any given time, we rank the importance of an agent or an expert based on its relations to other agents or experts. A CLA uses a time series model as follows:

- Data for a single agent that is observable, *e.g.*, a LLA model or an expert or expertise of a single agent is made from its stored content.
- Different types of expertise in different domains, a link matrix r_{ij} is used to describe how one expertise *i* is linked to another expertise *j*. Each agent or expert process its own data in parallel and separately.

An agent can have one or multiple expertise. For simplicity, we assume one agent only focuses on one of its best expertise. We also model the relation between agents or expertise using lexical links of a single agent's content input X_t and states as different types of expertise that are observed from the whole network, as a probability density function $b_j(X_t)$.

We use an Expectation and Maximization (EM) method to compute maximum likelihood estimates and compute the correlation or affinity between an input content *X* and a type of expertise *j* being implemented[28].

Let $b_j(X)$ be a likelihood function of producing content *X* if an agent possesses an expertise *j*. For a joint likelihood of multiple agents given all the parameters associated with a model λ

$f(X \mid \lambda) = \sum_{all\ s} \prod_{t=1}^{T-1} r_{s_{t-1}s_t} b_{s_t}(X_t)$, where *t=1, …, T* are samples. The results of learning λ is a recursion shown in Fig. 8 which maximizes a total reward *R(t,j)* of a multi-agent system up to time *t* with an ending expertise *j* including its connections r_{ij} to other agents, their total rewards *R(t-1,i)* up to *t-1* and agent *j*'s local expertise $b_j(X_t)$.

In Fig. 8, a system has two parts of reward it constantly tries to obtain: the authority reward and expertise reward. An authority reward measures how many connections exist between the system and a peer list (r_{ij}) of systems where the rewards of peers can be directly added to the system itself. The expertise reward or score $b_j(X_t)$ measures uniqueness of a system that possesses some information that no other systems have. The tradeoff between authority and expertise is controlled by the coefficients w_1 and w_2.

[27] Quantum Intelligence (QI) 2014. System and method for knowledge pattern search from networked agents. US patent 8,903,756

[28] Dempster A.P., Laird, N.M. and Rubin D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B 39: 1-38

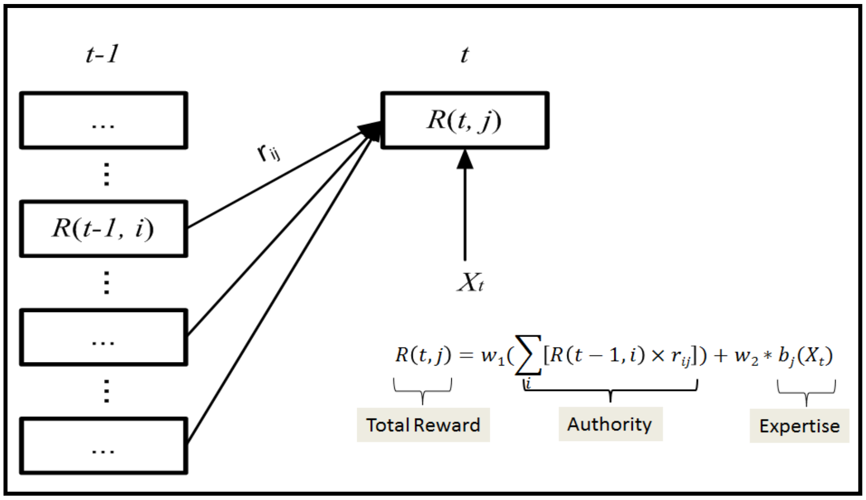

Fig. 8. Recursion to compute the overall reward of a system *R(t, j)*.

4.3 Relationship to Deep Learning

One important recent trend[29] is that some Big Data analytics such as Deep Learning methods including unsupervised machine learning techniques (*e.g.*, neural networks), sparse coding[30], and self-taught learning[31], have some interesting breakthroughs in that Big Data can be analyzed more intelligently and efficiently. For example, applied in machine vision and pattern recognition, self-taught learning [32] approximates the input for unlabeled objects as a succinct, higher-level feature representation of sparse linear combination of the bases. It uses the EM method to iteratively learn coefficients and bases[33]. Deep Learning links pattern recognition and text analysis intelligently. For example, LDA can be viewed as a sparse coding where a bag of words used as the sparsely coded features for text[34]. LLA uses bi-gram word pairs, compared to LDA, that are potentially more meaningful and sparse coded features. LLA is implemented in parallel processors, *i.e.*, each agent or expert analyzes its own data in parallel and results in a single global model as if all the data are processed together, *i.e.*, a sequence of expertise that optimize a global reward function.

In practice, one can argue that an imposter or spammer might adapt to fit the described features or profiles of authentic or authoritative archetypes, and can promote its own identity or page by simply copy and paste the text of top ranked profiles or pages into its own. In this case, the hyperlinks based ranking may be more successful because the links convey a notion of

[29] Reference 13

[30] Olshausen, B. and Field, D. 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature

[31] Raina, R., Battle, A., Lee, H., Packer, B., & Ng, A.Y. (2007). Self-taught learning: Transfer learning from unlabeled data. In ICML

[32] Ranzato, M., Monga, R., Devin, M., Chen, K., Corrado, G., Dean, J., Le, Q. V. and Ng, A.Y. Building high-level features using large scale unsupervised learning. Proceedings of the 29th International Conference on Machine Learning (ICML-12). Retrieved from http://arxiv.org/pdf/1112.6209v5.pdf

[33] LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. 1998. Gradient-Based Learning Applied to Document Recognition. Proceedings of the IEEE, 86(11):2278-2324. Retrieved from http://yann.lecun.com/exdb/publis/pdf/lecun-01a.pdf

[34] Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E. 2008. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment. 2008(10):P10008

endorsement by others, which is harder to manipulate in a large scale. However, when we deal with many Big Data without hyperlinks, applying a content-based authentication is a necessity. In this case, one can still use our approach by filtering out the popular themes, (*i.e.*, the ones that are already known or in the public conscience and awareness), then focus only on emerging and anomalous themes for sorting and ranking authority and authenticity. Emerging and anomalous features are more interesting in our use cases. For instance, the features that helped predict two persons with the same last name and first names belong to the “emerging” and “anomalous” categories.

5. Conclusions

While social media data mix tremendous diversified topics and personas, we showed the development of the algorithm, the collection and labeling of data, as well as the results of analysis of the characteristics of three authentic personas using a cyber professional as an example and the data collected from Facebook, LinkedIn and Twitter. The investigation of authentic archetypes and relatively anomalous themes containing the lowest number of mutually connected word pairs proved interesting to human analysts and provides insight with regard to demonstrated situational awareness. This method can inform automated sensemaking for analyzing and understanding massive Big Data for other applications.

At an operational level, this research is important for evaluating if a new persona (person or an organization) is authentic because of the following reasons: since archetypes of authentic personas are learned and discovered from Big Data and cross multiple public and social media platforms where we assume authentic or authoritative personas are dominant and therefore discovered and learned as archetypes, one could compare a new persona with the archetypes and generate a score that can be treated as a score of authenticity or authority.

Acknowledgment

We would like to thank Professor David Trask at the Naval Postgraduate School for his guidance. We also like to thank Dr. Charles Zhou from Quantum Intelligence, Inc. for his valuable input and discussion related to Collaborative Learning Agents (CLA).