Analyzing Multi-Dimensional Networks within MediaWikis

Brian C. Keegan
Northeastern University
Boston, MA, USA 02115
b.keegan@neu.edu

Arber Ceni
Social Media Research Foundation
Belmont, CA, USA 94002
arber@smrfoundation.org

Marc A. Smith
Social Media Research Foundation
Belmont, CA, USA 94002
marc@smrfoundation.org

ABSTRACT

The MediaWiki platform supports popular socio-technical systems such as Wikipedia as well as thousands of other wikis. This software encodes and records a variety of relationships about the content, history, and editors of its articles such as hyperlinks between articles, discussions among editors, and editing histories. These relationships can be analyzed using standard techniques from social network analysis, however, extracting relational data from Wikipedia has traditionally required specialized knowledge of its API, information retrieval, network analysis, and data visualization that has inhibited scholarly analysis. We present a software library called the `NodeXL MediaWiki Importer` that extracts a variety of relationships from the MediaWiki API and integrates with the popular NodeXL network analysis and visualization software. This library allows users to query and extract a variety of multidimensional relationships from any MediaWiki installation with a publicly-accessible API. We present a case study examining the similarities and differences between different relationships for the Wikipedia articles about "Pope Francis" and "Social media." We conclude by discussing the implications this library has for both theoretical and methodological research as well as community management and outline future work to expand the capabilities of the library.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

System

Keywords

Wikipedia, MediaWiki, NodeXL, network analysis, SNA, social media, visualization, data analysis

WikiSym '13, August 05–07 2013, Hong Kong, China
ACM 978-1-4503-1852-5/13/08 $15.00.

1. INTRODUCTION

`MediaWiki` is an open source wiki system used by Wikipedia as well as thousands of other wikis.[1] MediaWikis have a two-fold value: they not only contain rich content, but they also encodes a variety of rich relational meta-data about editors' contributions over time. The relationships within these meta-data can be extracted to reveal interactions that may support these large-scale collaborations. In aggregate, these connections form complex networks that can reveal structural patterns of the networks, key positions within them, as well as visualizations of sub-groups and the relationships between them. These maps allow networks to be contrasted, highlighting the structural properties of different collaborations or communities.

MediaWikis installations generally make much of their data publicly available in real time through an application programming interface (API). However, specifying the correct queries, retrieving the information, structuring and intersecting it with other data for analysis, and processing it for visualization are individually highly complex tasks requiring substantial knowledge across a variety of technical domains: users must be familiar with programming languages, information retrieval, data parsing and cleanup, graph data structures and manipulation, statistical analysis of graphs, and graph layout algorithms to be able to access, analyze, and visualize network data about wikis. We argue this complexity has, unfortunately, limited scholarly inquiry into these increasingly pervasive and influential information systems to only those scholars possessing these esoteric but nevertheless prerequisite skills. We introduce a software library that extracts these complex relationships using a simple graphical user interface and integrates with the popular `NodeXL` network analysis package.

After reviewing related work on network analyses of Wikipedia, we explain how to identify and extract multidimensional networks from MediaWikis. In particular, we describe a variety of relationships, article attributes, and editor attributes that can be retrieved from MediaWikis and describe the process of chaining relationships together for multidimensional network exploration. Then we introduce the design, implementation, and interface of the `NodeXL MediaWiki Importer` (NMWI). We use the NMWI in brief case studies to illustrate how the structural patterns of relationships vary significantly across different types of articles, editors, and categories on the English Wikipedia. We conclude by discussing the implications the NMWI has for scholarship in socio-technical systems.

[1] `http://www.mediawiki.org`

2. RELATED WORK

Wikipedia is an exemplar of a "knowledge network" in which concepts are explicitly linked to one another. However encyclopedias from Diderot's *Encyclopédie* through new conceptualizations of organizing knowledge proposed by Vannevar Bush, Ted Nelson, and Tim Bernders-Lee have employed the concept of relations to associate semantically-related topics together [29]. However, the scale, complexity, and openness of relationships in Wikipedia has made it a prominent corpus used predominantly by computer and information scientists but also increasingly by social scientists to understand large-scale, distributed collaboration.

Structural approaches to the analysis of social media data have theoretical importance for understanding how behavioral regularities and emergent social roles support and enable online community [9, 28]. New kinds of data large-scale and fine-grained data are enabling examination of behavioral change of entire social systems [20]. Prior work has analyzed the structure of editors contributing to articles [5, 18, 15], articles linking to other articles [14], editor modifying other editors' contributions [3, 13, 16], editors' discussions with other editor [19, 21], and changes in these structures over time [4, 16]. However, this work typically examines only a single relationship rather than multiple overlapping relationships.

3. OUR APPROACH

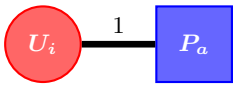
`NodeXL`[2] is an open-source network analysis application that uses the familiarity of Microsoft Excel spreadsheets to collect, store, analyze, visualize, and publish network datasets. `NodeXL` is focused on end user ease-of-use by simplifying the collection, analysis, and visualization of data from web and social media. Users can analyze data from Twitter, Facebook, YouTube, flickr, email, and the web without specialized knowledge about information retrieval, perform common network analysis tasks such as calculating centrality or analyzing community structure, and visualize the structure and attributes of the network [11, 25].

Although the English Wikipedia is by far the largest example of a MediaWiki-based online community,[3] MediaWikis are used in a variety of other contexts [26]. Because of the pervasiveness of this wiki system, we designed an importer that can analyze patterns of collaboration, communication, and interaction on Wikipedias in other languages [12], encyclopedia projects for fan and educational communities [8], and enterprises and distributed collaborations [10]. Because Wikipedia remains the most well-studied MediaWiki installation, we define relationships using Wikipedia as an example below and throughout the remainder of the paper. Despite this domain-specific definition, these relationships can be found and analyzed using the `NodeXL MediaWiki Importer` for any public MediaWiki API.

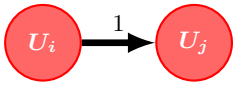
3.1 Relationship types

As the review above suggested, Wikipedia is a complex and multidimensional network that encodes a variety of relationships among editors, among articles, and between editors and articles. These networks are multidimensional because they not only encode a different of *types of nodes* but they also encode different *types of relationships*. Analysis of unidimensional networks often assumes distinct relationships can be collapsed together or altogether fails to capture the rich interdependencies among different sets of objects. Recognizing this plurality of network patterns and analyzing how these new and diverse structural signatures and their logics of affiliation are essential to understanding how these complex networks emerge, interact, and change [6]. We identify and review three broad classes of multidimensional networks within MediaWikis below.

(a) Undirected editor–article relationship with weight 1



(b) Directed editor–editor relationship with weight 1



(c) Directed article–article relationship with weight 3

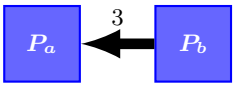


Figure 1: Examples of relationships (in black) involving editors (in red) and articles (in blue).

Networks fall into a variety of classes reflecting the types of nodes and types of edges in them [27]. *One-mode* (also known as "unimodal") networks have only one type of node in the network (*e.g.,* people) wherein links can exist between any pair of nodes in the network. *Two-mode* (also known as bipartite or affiliation) networks have two distinct types of nodes in the network (*e.g.,* people and objects) wherein links in the network can only exist between different types of nodes. Edges between nodes can be *directed* if the direction is meaningful and asymmetrical relationships can exist (*e.g.,* following on Twitter) or *undirected* if the direction is not meaningful and only symmetrical relationships can exist (*e.g.,* co-location). A variety of attributes can also be appended to both nodes and edges, which are described below.

3.1.1 Editor-article relationships

The creation of Wikipedia content requires editors to make revisions to a article. However, this relationship is inherently multimodal as it requires the representation of a relationship between two distinct types of nodes: editors (*e.g.*, editors) and articles (*e.g.*, articles). There is only a single substantive relationship in this class and an example is given in Figure 1(a).

Editing When a editor modifies a article this contribution represents an editing relationship. An editing link is created when editor i makes a change to article P. This is a two-mode network: a editor cannot modify another editor and a article cannot modify another article. This relationship can be weighted to reflect the number of times a editor modified a article within a window of time. It is traditionally undirected reflecting the lack of a meaningful "article modifying editors" tie. However, it is implemented as a directed tie with only one possible direction ("editor modifies article")

[2] `http://nodexl.codeplex.com`
[3] `http://meta.wikimedia.org/wiki/List_of_largest_wikis`

within NodeXL.

3.1.2 Editor-editor relationships

Wikipedia is also an online community wherein editors interact with other editors. These relationships are one-mode because they only involve editor nodes, but they can vary in their direction. There are three types of relationships in this class and an example is given in Figure 1(b).

Co-editorship This is a *projection* of the two-mode editing network into a one-mode editor-editor network. A link exists between editor i and editor j if both editors edited the same article. This is an undirected relationship because a editor cannot edit *at* another editor. This relationship can be weighted to reflect the number of articles two editors have both edited.

Discussions Writing wiki articles also demands explicit coordination by editors discussing the work related to a article or discussing work done by editors themselves. Threaded discussion articles exist for both editors and articles to develop consensus about how to write articles or respond to problematic editors [17, 19]. A link is created if editor i responds to editor j's comment on a discussion article. This is a directed relationship because one editor is responding the the other editor. This relationship can be weighted to reflect the number of times a editor responds to another editors.

Article trajectory The history of changes made to a article can be associated with individual editors such that every change can be interpreted as one editor modifying a version of a article previous made by another editors [16]. Alternatively, this can be interpreted as a "document passing" network tracing the history of editors' changes. For a given article, a link exists from editor i to editor j if editor j modified the article following editor i. The resulting network of editor-editor interactions is unique to every article. This is a directed relationship because one editor modifies another editor's version of the article. This relationship can be weighted because one editor can repeatedly modify another editor's versions of a article.

3.1.3 Article-article relationships

The previous relationships encode information related to the editors' contributions to articles, but the articles themselves also encode relationships based on their content. There are four types of relationships in this class and an example is given in Figure 1(c).

Shared editorship This is a *projection* of the two-mode editing network into a one-mode article-article network. A link exists between article A and article B if both articles were modified by a common editor. This is an undirected relationship because a article cannot edit *at* another article. This relationship can be weighted to reflect the number of editors two articles shared in common.

Hyperlinks Wikipedia articles link to other Wikipedia articles in their text. A hyperlink relationship exists if article A has a wiki-link to article B. These links are directed because hyperlinks are not symmetric: a article may link to another without the other linking back.

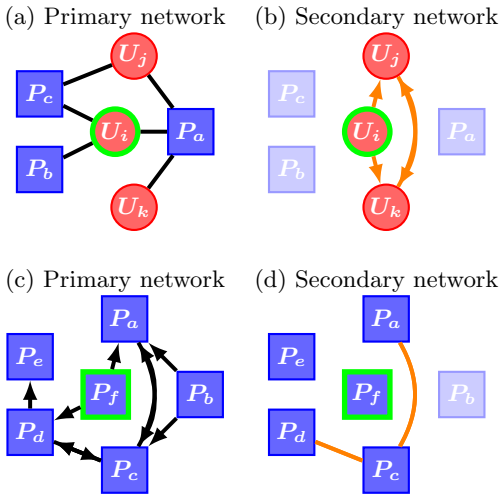


Figure 2: Illustration of (a) an undirected two-mode primary network and its (b) hypothetical directed one-mode secondary network. (c) is a directed one-mode primary network and (d) is its hypothetical undirected one-mode secondary network. The "ego" is outlined in green.

These relationships are unweighted to reflect only that a link exists rather than the number of times a article is linked.

Editor trajectory The history of a editor's contributions can be interpreted as a editor moving from one article to another. For a given editor, a link exists from article A to article B if article B was modified by the editor after he or she modified article A. The resulting network of article-article interactions is unique to every editor. This relationship is directed because a editor moves from article A to article B. This relationship can be weighted because a editor can repeatedly move from article A to article B.

Category co-membership Articles on Wikipedia are classified into topical categories with other articles. If article A and article B are members of the same category X, they will share a link. This relationship is undirected because the relationship is symmetrical: a article can't have a one-way co-membership link. The relationship is weighted and reflects the number of categories both articles share membership. Substantively, pairs of articles that share many category co-memberships are likely to be strongly related.

3.2 Multidimensional network exploration

To generate a set of nodes ("alters") from a seed node ("ego"), we use the relationships defined above. However, this approach can be extended to explore the multidimensionality of the network by *chaining* two different types of relationships together. Given a seed node and a *primary relationship*, NMWI will return all the alters to which the ego is connected. The vast majority of network analyses stop here and NMWI editors can as well. However, the set of nodes consisting of the ego and its alters (or its alters' alters) can then be examined for *secondary relationships*. In effect, overlapping relationships can be parsimoniously explored by extracting a set of nodes based on a primary relationship but

then examining if there are secondary relationships within this set of nodes.

For example, specifying an editing relationship as the primary relationship for a seed editor will return a list of all the articles who have been edited by the seed editor. If the ego in Figure 2(a) is editor U_i, her alters can only be articles because the network is two-mode. The alters of U_i are articles P_a, P_b, P_c and her alters' alters are editors U_j and U_k. The ego editor and her alters' alter editors in this two-mode primary network are a set of editors (U_i, U_j, U_k) joined by the relationship of contributing to the same articles. The multidimensionality of this relationship can be illuminated by using the set of editors defined in the primary relationship as seeds for a secondary relationship: do these editors communicate with each other? Specifying discussions as the secondary relationship, Figure 2(b) shows the secondary discussion relationships among editors *conditional on* these editors appearing in the primary editing relationship while P_a, P_b, P_c are excluded for being a different type. Chaining the primary and secondary relationships together reveals an overlapping network in which editors co-editing articles together also have discussion relationships with each other.

In the other example, if article P_f is the ego and the relationship is hyperlinks, its alters in Figure 2(c) are also articles because the network is one-mode, but link direction in a directed network must be respected. The alters of P_f are articles P_d and P_a and its alters' alters are articles P_c and P_e but *not* P_b because P_f cannot reach P_b following link directions. The multidimensionality of this relationship can be examined by chaining the set of articles defined in the primary relationship to a secondary relationship such as category co-membership. Figure 2(d) shows the secondary category co-membership relations on the set of articles identified from the primary relationship while P_b is excluded for not being in the primary set. This multidimensional analysis suggests articles with reciprocal hyperlinks (the primary relationship) also have category co-membership relations (the secondary relationship). These overlapping relations are explored in the case study in Section 5.

4. SYSTEM IMPLEMENTATION

The MediaWiki API provides direct access to a variety of data contained within its databases.[4,5] The revision histories of individual articles can be extracted to create editing networks, the revision histories of editors can be extracted to created editor trajectory networks, the links to and from a article can be queried to create a hyperlink network, and the content of talk articles can returned to create discussion networks. Data for the English Wikipedia (the largest and oldest MediaWiki installation) goes back to mid-2002, but the availability of historical data for other MediaWikis may vary based on installation and administration practices.

The `NodeXL MediaWiki Importer`[6] (NMWI) is a graph data provider that allows users to download data from public MediaWiki APIs and import them into `NodeXL` [11]. It is developed in C# and employs the `DotNetWikiBot` framework for interfacing with a MediaWiki API.[7]

[4] `http://www.mediawiki.org/wiki/API:Main_page`
[5] `http://en.wikipedia.org/w/api.php`
[6] `https://wikiimporter.codeplex.com/`
[7] `http://dotnetwikibot.sourceforge.net/`

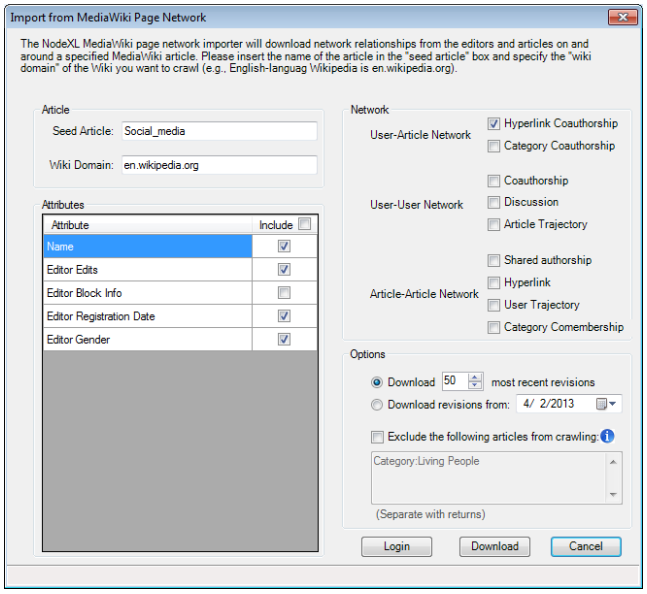


Figure 3: NodeXL MediaWiki Importer interface.

As a wrapper around the API, the NMWI simplifies the extraction, cleaning, and structuring MediaWiki API output for the `NodeXL` end-user. This framework wraps the main API calls in three main classes, respectively `Site`, `Page` and `PageList`. The `Site` class represents a MediaWiki site (*i.e.*, `http://en.wikipedia.org`) and implements methods to gather general information about the site. The `Page` class is the main class which represents a MediaWiki article which allows us to load the article content, links, categories it belongs to, images and other information related to an article. If we want to search for articles, load articles from a category or get all the articles linked by a seed articles we have to make use of the `PageList` class. Although `DotNetWikiBot` implements most of the functionality needed to parse a MediaWiki site, we introduced a new function to accommodate the download of revisions within specific time ranges and date as well. Furthermore, because `NodeXL` can use images as vertex representations, we also download the image URLs for alternative representations.

To create the different networks available in NMWI we download a list of articles which depending on the choice is a list of all hyperlinks in the seed article or the articles of the categories that the seed article belongs to. Then for each of this article names we download a list of its revisions, including the editor information, depending on the criteria specified. These data are stored in a dictionary keyed by article name and with values corresponding to the list of revision information such as editor name, size, and comment. Subsequent network structures can be created by traversing this data structure.

4.1 Interface

The NMWI is installed separately from `NodeXL` and provides a new option for importing data. When launched, the user is presented with the Importer interface as seen in Figure 3. Users perform several steps to identify the seed entity from which data will be retrieved, specify the types of alters to be extracted, define the types of relationships to be

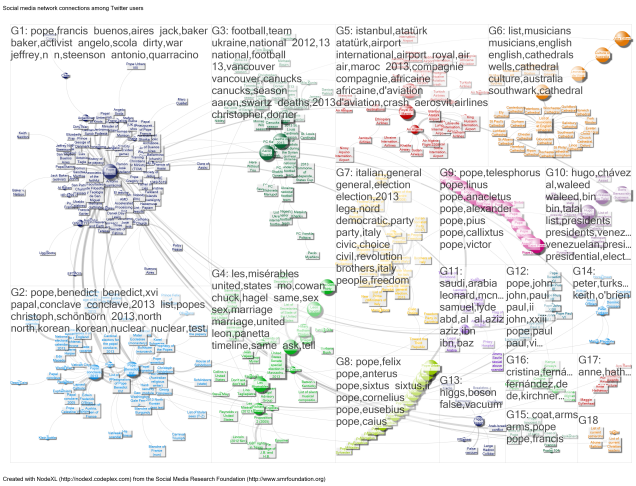


Figure 4: Two-mode editing network for Pope Francis with clusters arranged into boxes and labeled by article titles.

returned, and limit the query by certain parameters.

First, the user enters the name and type of a seed entity (article, editor, category, or file) and domain for the MediaWiki installation from which the importer will extract data. Some types of alters or relationships in subsequent steps cannot be extracted for a given seed entity and they are greyed out. Second, given this seed entity, the user selects a primary relationship for identifying "alter" nodes . Once the set of alters has been identified, the primary relationships among these alters can be analyzed or, optionally, can be used as an input to analyze a secondary relationship (see Section 3.2).

Because editors and articles can potentially have hundreds of revisions, editors, links, or category members, NMWI specifies options to return only the *most recent* revisions or revisions within a *time window*. After extracting the data, the NMWI populates the "Nodes" and "Edges" tabs of the Excel worksheet.

5. CASE STUDY

The method for chaining network relations together to explore the multidimensional character of these collaborations provides many potential permutations. We explore only six relationships below based on both article-article relationships and editor-editor relationships. To demonstrate the capabilities of the NMWI for large-scale and multidimensional network analysis, we compare the Wikipedia articles "Social media"[8] and "Pope Francis"[9] from late March 2013 for revisions made since January 1, 2013. The contrast between these articles should be instructive as former attracts consistent and sustained attention while the latter is emblematic of an article that attracts sudden and intense attention. As such, we expect to see significant differences in the structures of the various networks extracted from NMWI.

The networks visualized in Figures 5–9 are plotted using the "Harel-Koren Fast Multiscale" layout and the nodes are grouped and colored by "Clauset-Newman-Moore" clustering. It is important to note that the same node may appear within different groups depending on the relationship and thus its color and position may vary across networks. Inter-group edges are combined to show the density of ties between groups while also keeping within-group structure clear. Editor-editor networks for co-editorship, trajectories, and discussion and article-article networks for shared editorship and hyperlinks are discussed below.

5.1 Editor-editor networks

5.1.1 *Shared article editing networks*

The co-editorship networks in Figure 5 are projections of the two-mode editor-article networks for all the revisions made by every editor contributing since January 1, 2013 and the other articles they edited during this window. A link exists from one editor to another if they both edited the same article during this time and the weight of this link corresponds to the number of articles they jointly edited. Because all the editors revised at least the seed article in common, these networks reflect edges of weight two or more. Both networks exhibit strong clustering and the identified subgroups also have strong ties to the other subgroups as editors jointly revise many articles together.

5.1.2 *Article trajectories*

The article trajectory networks in Figure 6 capture the which editor subsequently changed an article after another editor. These trajectories encode information about the temporal context (when a editor made contributions relative to other editors also making contributions) as well as editor engagement (editors contributing repeatedly and in response to diverse editors) [16]. The clusters in Figure 6(a) are both larger and more dense than the clusters observed in Figure 6(b), which is unsurprising as editors contributing to a article about a current event are likely more engaged and motivated to edit repeatedly as they update information and respond to others' contributions. In contrast, the "spaghetti" subgraph seen in the top-center box of Figure 6(b) reveals a subgroup of editors making only a single contribution and never returning to edit again. Similarly, the editor at the middle-center of the cluster in the central box in Figure 6(b) is actually a bot doing automated work, not a human editor. This structural signature of these editing patterns can serve a useful diagnostic function as they reveal a lack of editor engagement in contrast with the highly-clustered patterns in Figure 6(a) indicative of sustained contributions.

5.1.3 *Discussion networks*

The discussion networks in Figure 7 capture whether editors who contributed to either of these articles also interact with each other on their respective editor talk articles. Despite the differences in the other relationships explored above, both articles exhibit similar structural patterns and overall size. In fact, the same two editors appear as the most central editors in *both* networks: "Arctic Kangaroo"[10] and "Materialscientist"[11] are prolific recent-change patrollers[12]

[8] http://en.wikipedia.org/w/index.php?title=Social_media&oldid=547781893

[9] http://en.wikipedia.org/w/index.php?title=Pope_Francis&oldid=547771007

[10] http://en.wikipedia.org/wiki/User:Arctic_Kangaroo

[11] http://en.wikipedia.org/wiki/User:Materialscientist

[12] http://en.wikipedia.org/wiki/Wikipedia:Recent_changes_patrol

(a) Pope Francis

(b) Social media

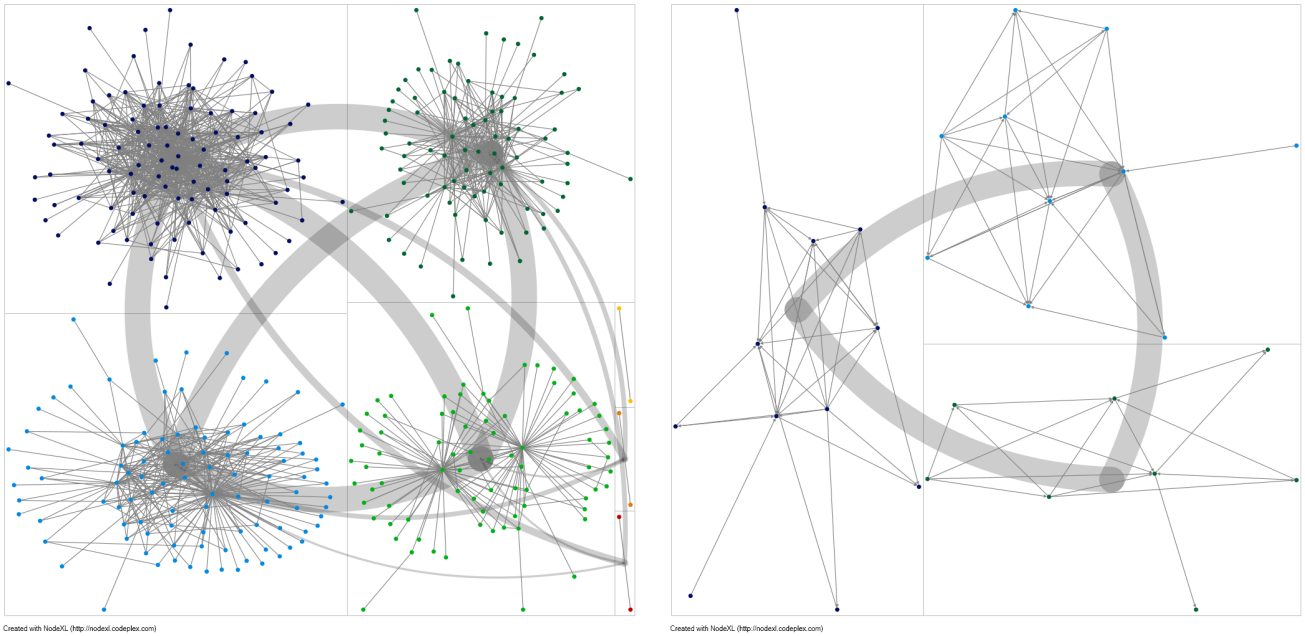


Figure 5: Undirected editor-editor shared editorship networks.

(a) Pope Francis

(b) Social media

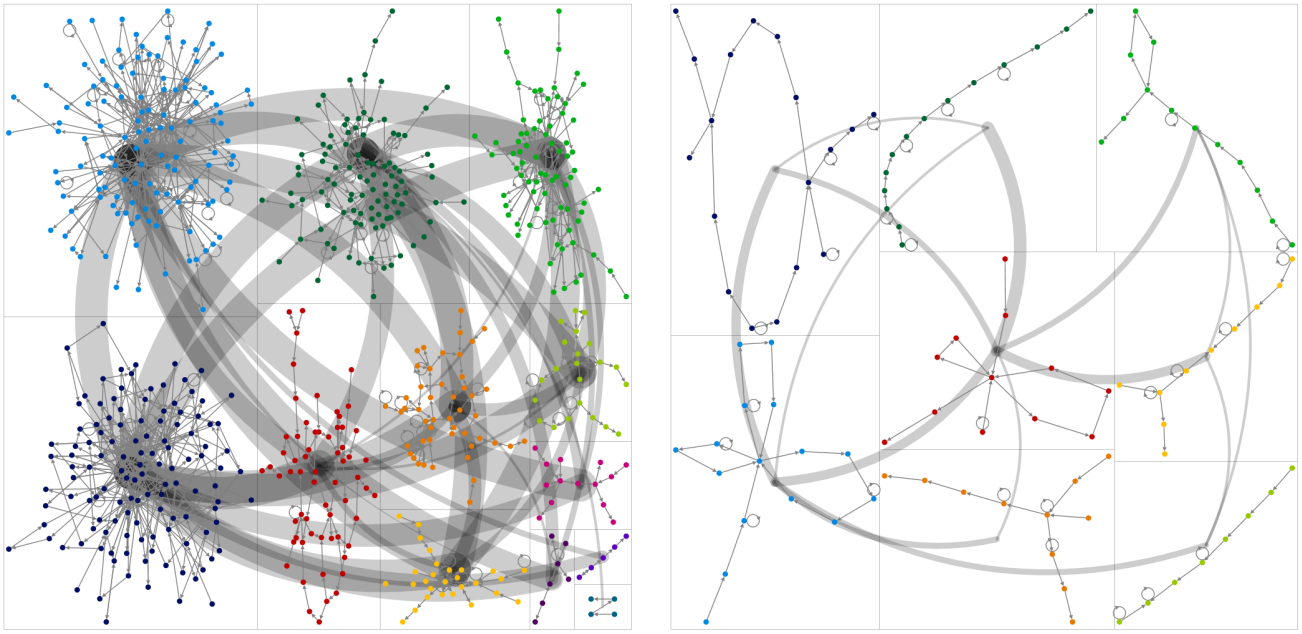


Figure 6: Directed editor-editor article trajectory networks.

who monitor articles for vandalism. Their centrality in both articles reflects first, the revisions they made to both articles reverting vandalism and second, their leaving messages on the responsible editors' talk articles informing or warning them. These editors fulfill crucial social roles on Wikipedia and the discussion network vivdly captures an important dimension of their work trying to keep bad content out while trying to bring new editors in.

5.2 Article-article networks

5.2.1 Shared editor editing networks

The co-editorship networks in Figure 8 are projections of the two-mode editor-article networks for all the revisions made by every editor contributing to each article between January 1, 2013 and March 30, 2013 and every other article these editors revised over the same period of time. Given

(a) Pope Francis

(b) Social media

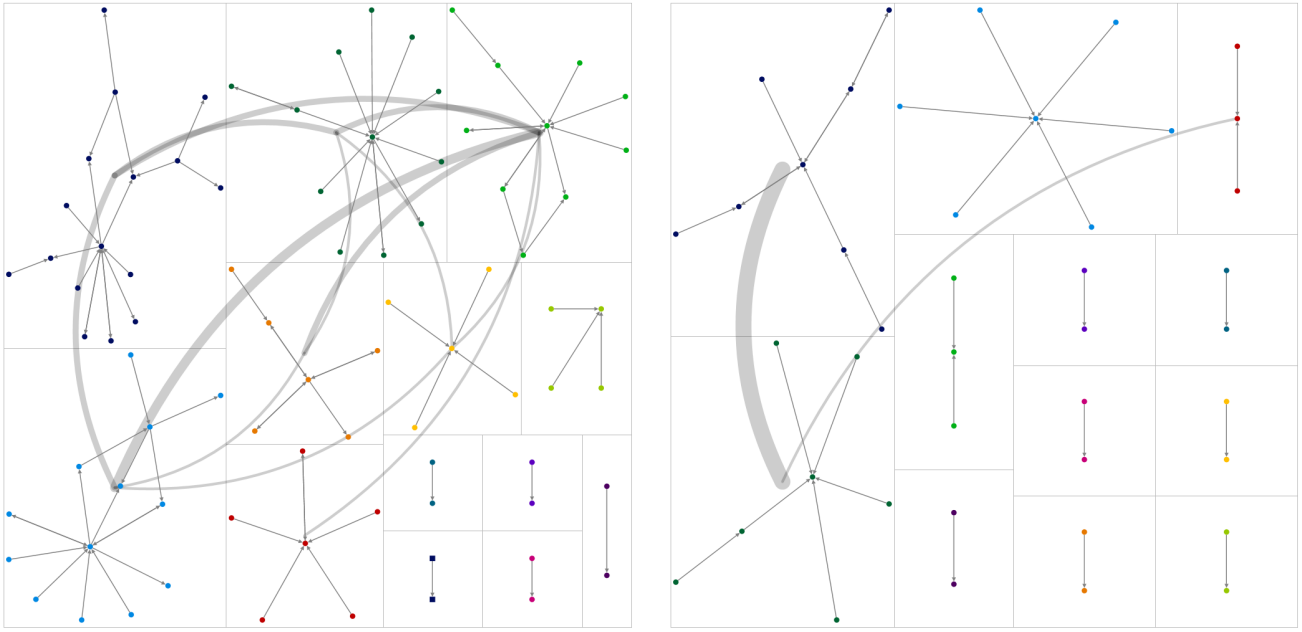


Figure 7: Directed editor-editor discussion networks.

(a) Pope Francis

(b) Social media

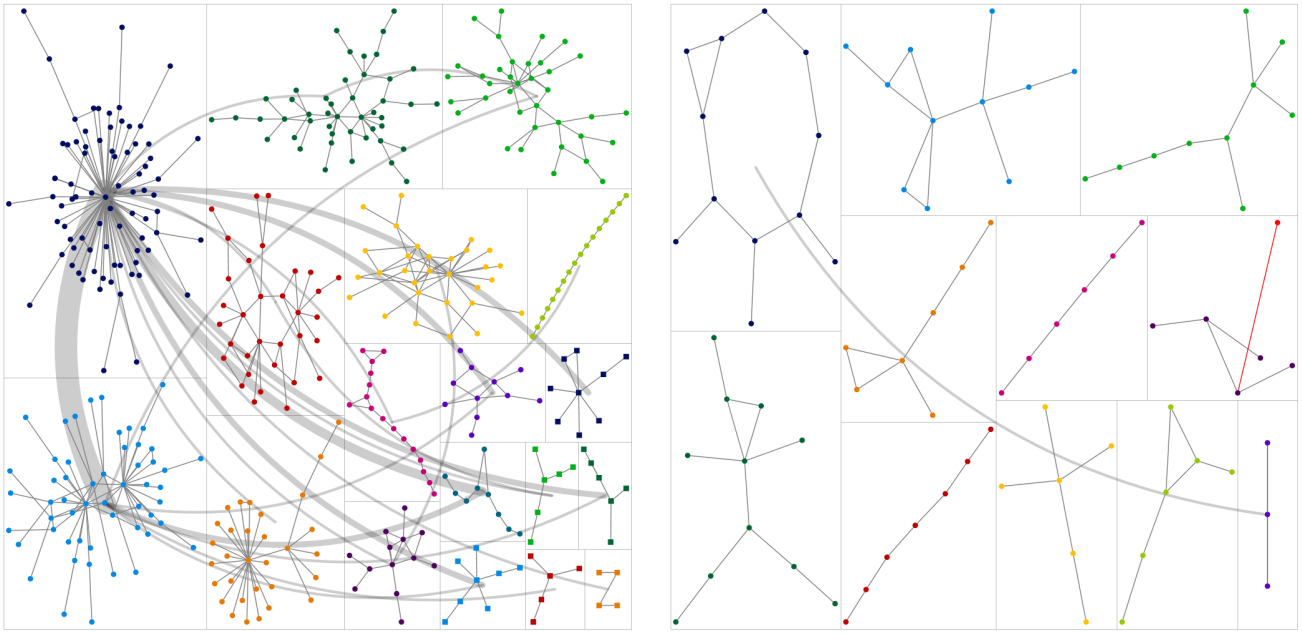


Figure 8: Undirected article-article co-authorship networks.

the raw size of the resulting networks, edges were filtered out if articles only shared a single editor in common and the remaining largest connected component was visualized.

As expected, the current event article has a substantially larger co-editorship network than the traditional article. Most of the other clusters in Figure 8(a) correspond to topics related to the papal transition or other current news events such as the Italian election and Academy Awards suggesting there are many Wikipedia editors engaged in newswork both within and across topics. The clusters in Figure 8(b) for "Social media" are less dense and have fewer between-group connections. The groups themselves correspond to esoteric topics about cartoons and haplogroups that bear little topical similarity to social media and reflect the diverse interests (and potential biases) of the editors contributing to it.

5.2.2 *Hyperlink networks*

The directed hyperlink networks capture the alter articles to which the seed article links as well as whether these alter articles link to one another. Figure 9 shows a marked difference in the structures of the local hyperlink networks for each article. The links on Pope Francis are extremely dense as a result of a navigation templates. This is troublesome as these links are redundant and not derived from the body of the text, but rather from a box of links at the bottom of each article linking to every other article about a Roman Catholic pope. Future iterations of NMWI will instead parse the content of the article itself to extract only the links appearing in the text rather all links present anywhere on the article. The cluster of articles on the right corresponds to articles about every Pope and the history of the Catholic Church while the left cluster shows less, but still very dense, links around other topics from the text.

This stands in marked contrast to the low level of clustering observed among the "Social media" articles. The cluster on the left in Figure 9(b) corresponds to a variety of articles about editors and books and "newer" social media topics like crowdsourcing, the cluster on the lower right to articles about social networking services, and the cluster on the upper right to "older" social media topics topics like blogging and podcasting. However, the lack of completely-connected clusters is also revealing as it indicates that the editors editing these articles potentially lack the commitment to create a similar box to link the same topics together across articles or that the articles themselves do not lend themselves to being grouped together into a common topic.

6. DISCUSSION

The `NodeXL MediaWiki Importer` is a tool for analyzing the multidimensional network relationships embedded within `MediaWiki` installations such as Wikipedia. The tool enables end users to perform complex data collection, analysis, visualization, and publication tasks with a graphical interface accessible to non-experts. This tool adds large-scale knowledge and collaboration networks as another type of graph data provider for `NodeXL` in addition to existing ones such as Twitter, Facebook, e-mail, and the web. The NMWI uses publicly-accessible and lightweight API calls for data extraction from the live versions of articles for all public MediaWikis rather than the heavyweight and quickly-outdated database dumps released only by the Wikimedia Foundation. As such, this tool expands the repertoire of real time social media analytics that have primarily focused on Twitter: Wikipedia's and any other MediaWiki's updates about breaking news can be analyzed in real time.

Our case study used NMWI to examine two articles on the English Wikipedia: "Pope Francis" was a major current event that attracted intense activity from hundreds of editors and "Social media" was an article emblematic of more traditional patterns of sustained collaboration from fewer editors. We found striking similarities and stark differences across several dimensions of relationships within these articles. The case only explored a fraction of the permutations of possible multidimensional relations that could be examined using the chaining approach described in Section 3.2. For example, co-editorship patterns could be examined among the alters generated by hyperlinks or category co-membership primary relationships rather than the editing patterns examined here.

This tool has the potential for advancing methods in social network analysis and theories of collective behavior as well as supporting more applied interests such as diagnosing the health of an online community. Because the organizing logic of one network may be predicated on the presence or absence of ties in other relationships, data about these covariate networks must also be collected to develop appropriate statistical models for inference [22]. Theories of social roles and repertoires of action in online peer production communities have traditionally examined regular patterns of interaction articulated through a single relationship [9, 28]. The multidimensional data identified by the NMWI offers the possibility of identifying richer and more nuanced roles. Finally, wiki community managers can use NMWI to identify key participants or examine the effect of interventions to improve collaboration without the need for specialized analysis.

Additional features are worth considering in future versions of the NMWI such as the attributes of editors and articles, the dynamic features of relationships, employing the content of articles, and examining additional relationships. First, articles and editors are not homogeneous and further work should be done to create and import rich attributes for the nodes in the network to compliment the richness of the relationships. A variety of attributes for editors can be extracted such as the editor type to differentiate unregistered editors from registered editors and admins, the total number of edits made by editors as a proxy for their expertise, the total number of blocks imposed to understand how controversial they are, their tenure within the community, or how concentrated is their activity in some articles. Analogously, attributes for articles can also be extracted such as the type of article to differentiate articles from discussion articles, the age and edit counts as a proxy for quality, and the concentration of editing from some editors. Attributes are essential as they can both *influence* and *be influenced by* the structure of links [24].

Second, networks are not static and additional features can be implemented to record timestamped information about editors, articles, and the relationships between them to understand their evolution. Contributions to articles may be highly uneven and "bursty", editors may exhibit cyclical patterns in their contribution activity, links can be introduced or removed, and there are rich time series data available for article view activity across the Wikimedia Foundation projects. Understanding how these dynamics play out are crucial for establishing causality but also introduce significant challenges for efficient data storage and parsimonious data visualization. Third, the content of articles themselves can be more fully incorporated for network text analysis and content analysis [7]. Prior work has examined the persistence of content within Wikipedia articles [23, 1] suggesting that editing patterns could be examined in finer detail around particular sections or types of content being introduced and removed. Finally, content across language editions reveal fascinating similarities and differences in how topics are constructed based on the patterns of their links [2]. These inter-language relationships as well as patterns in the external sites Wikipedia and other MediaWikis link to can provide rich information about the larger information ecosystem in which wikis play such a central role.

(a) Pope Francis

(b) Social media

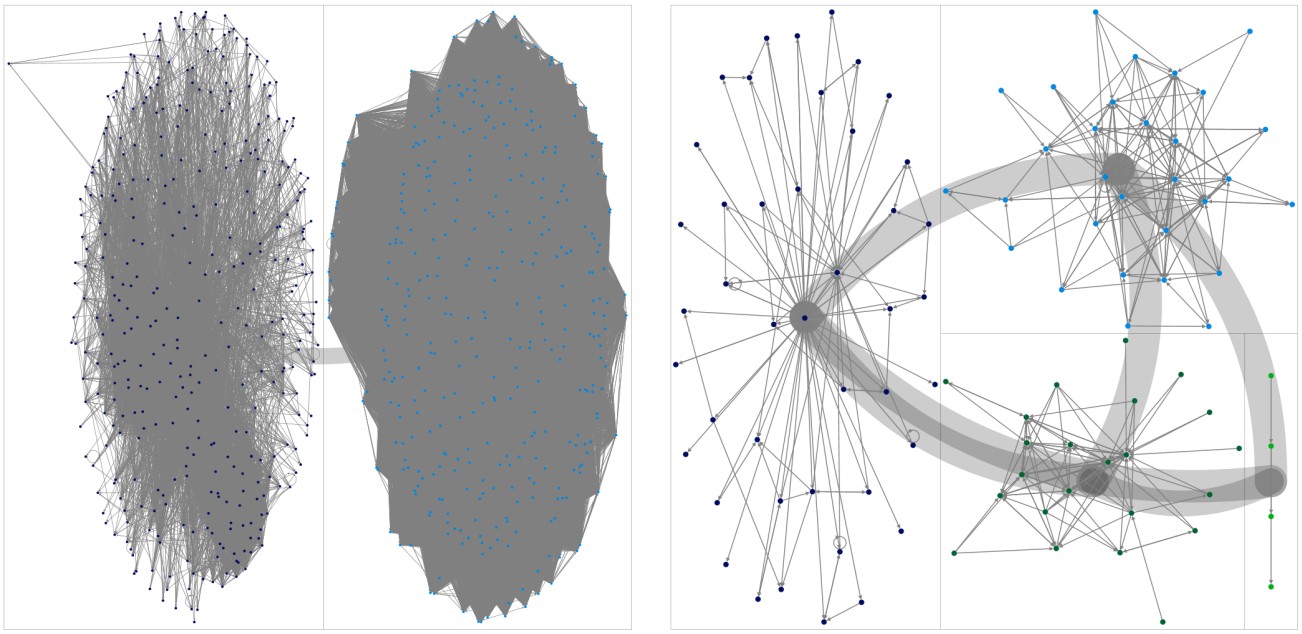


Figure 9: Directed article-article hyperlink networks.

7. CONCLUSIONS

We developed the `NodeXL MediaWiki Importer` to extract complex multidimensional relationships from the `MediaWiki` API for analysis by non-expert users. The NMWI provides a graphical user interface that builds upon the `NodeXL` platform to extract, analyze, visualize and publish networks from Wikipedia as well as thousands of other public MediaWiki installations.

8. ACKNOWLEDGMENTS

We would like to thank the users of the `NodeXL` application and the supporters of the Social Media Research Foundation.

9. REFERENCES

[1] J. Antin, C. Cheshire, and O. Nov. Technology-mediated contributions: editing behaviors among new wikipedians. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, CSCW '12, pages 373–382, New York, NY, USA, 2012. ACM.

[2] P. Bao, B. Hecht, S. Carton, M. Quaderi, M. Horn, and D. Gergle. Omnipedia: bridging the wikipedia language gap. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, CHI '12, pages 1075–1084, New York, NY, USA, 2012. ACM.

[3] U. Brandes, P. Kenis, J. Lerner, and D. van Raaij. Network analysis of collaboration structure in wikipedia. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 731–740, New York, NY, USA, 2009. ACM.

[4] L. S. Buriol, C. Castillo, D. Donato, S. Leonardi, and S. Millozzi. Temporal analysis of the wikigraph. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 45–51. IEEE, 2006.

[5] A. Capocci, V. D. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli. Preferential attachment in the growth of social networks: The internet encyclopedia wikipedia. *Physical Review E*, 74(3):036116, 2006.

[6] N. Contractor, P. R. Monge, and P. Leonardi. Multidimensional networks and the dynamics of sociomateriality: Bringing technology inside the network. *International Journal of Communication*, 5:682–720, 2011.

[7] J. Diesner and K. M. Carley. Revealing social structure from texts: meta-matrix text analysis as a novel method for network text analysis. In *Causal mapping for information systems and technology research: Approaches, advances, and illustrations*, pages 81–108. Harrisburg, PA: Idea Group Publishing, 2005.

[8] A. Forte and A. Bruckman. From wikipedia to the classroom: exploring online publication and learning. In *Proceedings of the 7th international conference on Learning sciences*, ICLS '06, pages 182–188. International Society of the Learning Sciences, 2006.

[9] E. Gleave, H. T. Welser, T. M. Lento, and M. A. Smith. A conceptual and operational definition of 'social role' in online community. In *Proceedings of the 42nd Hawaii International Conference on System Sciences*, HICSS'09, pages 1–11. IEEE, 2009.

[10] J. Grudin and E. S. Poole. Wikis at work: success factors and challenges for sustainability of enterprise wikis. In *Proceedings of the 6th International Symposium on Wikis and Open Collaboration*, WikiSym '10, pages 5:1–5:8, New York, NY, USA, 2010. ACM.

[11] D. Hansen, B. Shneiderman, and M. A. Smith.

Analyzing social media networks with NodeXL. Morgan Kaufmann, 2010.

[12] B. Hecht and D. Gergle. The tower of babel meets web 2.0: user-generated content and its applications in a multilingual context. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 291–300, New York, NY, USA, 2010. ACM.

[13] R. Jesus, M. Schwartz, and S. Lehmann. Bipartite networks of wikipedia's articles and authors: a meso-level approach. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, page 5. ACM, 2009.

[14] J. Kamps and M. Koolen. Is wikipedia link structure different? In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 232–241, New York, NY, USA, 2009. ACM.

[15] B. Keegan, D. Gergle, and N. Contractor. Do editors or articles drive collaboration?: multilevel statistical network analysis of Wikipedia coauthorship. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 427–436. ACM, 2012.

[16] B. Keegan, D. Gergle, and N. Contractor. Staying in the loop: Structure and dynamics of Wikipedia's breaking news collaborations. In *Proceedings of the 8th International Symposium on Wikis and Open Collaboration*. ACM, 2012.

[17] A. Kittur and R. E. Kraut. Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, CSCW '08, pages 37–46, New York, NY, USA, 2008. ACM.

[18] D. Laniado and R. Tasso. Co-authorship 2.0: patterns of collaboration in wikipedia. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia*, HT '11, pages 201–210, New York, NY, USA, 2011. ACM.

[19] D. Laniado, R. Tasso, Y. Volkovich, and A. Kaltenbrunner. When the wikipedians talk: Network and tree structure of wikipedia discussion pages. *Proceedings of ICWSM*, 2011.

[20] D. Lazer, A. S. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, et al. Life in the network: the coming age of computational social science. *Science*, 323(5915):721, 2009.

[21] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Signed networks in social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1361–1370, New York, NY, USA, 2010. ACM.

[22] P. R. Monge and N. S. Contractor. *Theories of communication networks*. Oxford University Press, 2003.

[23] R. Priedhorsky, J. Chen, S. T. K. Lam, K. Panciera, L. Terveen, and J. Riedl. Creating, destroying, and restoring value in wikipedia. In *Proceedings of the 2007 international ACM conference on Supporting group work*, GROUP '07, pages 259–268, New York, NY, USA, 2007. ACM.

[24] C. R. Shalizi and A. C. Thomas. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research*, 40(2):211–239, 2011.

[25] M. A. Smith, B. Shneiderman, N. Milic-Frayling, E. Mendes Rodrigues, V. Barash, C. Dunne, T. Capone, A. Perer, and E. Gleave. Analyzing (social media) networks with nodexl. In *Proceedings of the fourth international conference on Communities and technologies*, pages 255–264. ACM, 2009.

[26] J. Stuckman and J. Purtilo. Measuring the wikisphere. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, WikiSym '09, pages 11:1–11:8, New York, NY, USA, 2009. ACM.

[27] S. Wasserman and K. Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[28] H. T. Welser, D. Cosley, G. Kossinets, A. Lin, F. Dokshin, G. Gay, and M. Smith. Finding social roles in wikipedia. In *Proceedings of the 2011 iConference*, pages 122–129. ACM, 2011.

[29] M. Zimmer. Renvois of the past, present and future: hyperlinks and the structuring of knowledge from the encyclopédie to web 2.0. *New Media & Society*, 11(1-2):95–113, 2009.