Literary analysis and complex networks

RANZIVELLE MARIANNE L. ROXAS-VILLANUEVA

Our world is made up of multiple interconnected agents. The interactions between these components can be represented using complex networks. A network is composed of nodes, that represent the agents, and edges which are links that indicate interactions, connections or relationships between the nodes. For example, in the Facebook social network the people are the nodes and the edges denote friendship. Other examples of complex networks are neural networks, metabolic networks, food webs, transportation networks, the World Wide Web, the Internet, social networks of connections between individuals, organizational networks and networks of business relations between companies, networks of citations between papers, and many others. Some examples of real world networks are presented in Table 1.

Networks	Node	Link
financial network	banks	business transactions
transportation network	airport	airline route
neural network	neurons	synapses
metabolic networks	enzymes and metabolites	metabolic reactions
food web	species	predator-prey relationship
Facebook/ Twitter social network	individuals	“friendship”/”following”
semantic network	word	semantic relationship (eg.synonymy,

Table 1: Some examples of real world networks.

To quantify the structure of a network, statistical measures can be calculated. Parameters that are usually used to describe complex networks are summarized in Table 2. BC is the fraction of all shortest paths in the network that pass through a given node. A high value of the normalized BC means that a large number of nodes are connected to each other by short paths. The average C would give us an idea of how well connected the local neighborhood of a node is. If its neighbors are fully connected to each other, then C = 1 while a value close to 0 means that there are few connections within the neighborhood. The mean path length D is the average distance between any two vertices and the average diameter is the maximum distance between any two nodes in the network.

Network Parameter	Description
Network size (N)	total no. of nodes in the network
Average degree (D)	average number of edges per node
Mean path length (L)	$L = \frac{2}{N(N+1)} \sum_{i \geq j} d_{ij}$ where *dij* is the smallest number of links from node *i* to *j*
Average diameter (AD)	the longest path or largest no. of links connecting any pair of nodes in the network
Average clustering coefficient (C)	$C = \frac{1}{N} \sum_{i \in V} \frac{2l_i}{k_i(k_i - 1)}$ where *V* is the set of all nodes in the network and k_i is the number of neighbors or nodes directly connected to node *i*
Normalized betweenness centrality (BC)	$BC = \frac{1}{(N-1)(N-2)} \sum_{i \neq v \neq j \in V} \frac{\sigma_{ij}(v)}{\sigma_{ij}}$ where σ_{ij} is the number of *dij*'s and $\sigma_{ij}(v)$ is the number of paths from node *i* to node *j* that include node *v*

Table 2. Network metrics

Most real world networks exhibit small-world and scale-free features. Small-world networks are found to have a short characteristic path length and high clustering coefficient (Watts and Strogatz 440). That is, no matter how large the network is, any two nodes in the network can be connected through a small number of intermediate nodes. Nodes are likely to share common neighbors and form clusters. Scale-free networks, on the other hand, exhibit a power- law degree distribution which indicates the existence of some highly-connected nodes (hubs) (Barabási and Albert 509).

There has been a rapid growth of interest in studying networks in various disciplines. Many real-world systems, which had not been studied as networks, are now analyzed in networks perspective. In epidemics studies, for example, when the model takes into account the population structure as a scale-free network, it was shown that a more effective way of stopping epidemics is through identifying and immunizing the hubs in the network (Pastor-Satorras and Vespignani 036104).

With these advances in other field of research, interests in networks related to language and text analysis are growing as well. Typical network studies in computational linguistics have nodes to represent words and edges the syntactic or semantic relationship between them. One study constructed a network of synonyms using Moby thesaurus dictionary with links between synonymous words (Motter et al. 065102). The network shows small-world features; it is highly-clustered and has a small path length. Similar findings, including degree distribution following a power law, were obtained from three networks built from different sources: one is based on a free-association database, one uses data in WordNet, and a synonym network using Roget's thesaurus (Steyvers and Tenenbaum 41). Hierarchical structures from networks built from text have also been found to follow long-range correlations that reflect the changes in content within it (Alvarez-Lacalle et al. 7956). A network of lexical collocations has also been used to segment texts into thematically coherent units (Ferret 1481).

Another way of constructing networks is to link words according to their collocation or grammatical relationships. Ferrer and Solé linked two words if they are adjacent neighbors in a sentence (Ferrer i Cancho and Solé 2261). They used the British National Corpus for network construction. There are also studies on the properties of collocation networks for other languages, like Russian (Kapustin and Jamsen 89), Chinese (Liang et al. 4901) and many others (Gao et al. 579). The basic topological properties of the networks, small-world and scale-free, are similar across languages which indicate that these characteristics are linguistic universals like Zipf's law.

Network approach was also used to analyze individual texting styles and the information change of SMS text messages in Filipino (Cabatbat and Tapang 9). Analyzing complex networks can also reveal patterns such as news frames in texts/ articles to understand how the public position themselves on different sides of arguments like those involving population and family planning issues in the Philippines. Results show that disagreement on what are suitable population policies is due to the mismatched frames within which the issues are discussed (Legara, Monterola, and David 4600; David et al. 329; Legara et al. 51).

In the following sections, we present some of the existing works done in the Philippines on the application of complex network approach in the analysis of literary pieces.

QUANTIFYING THE DIFFERENCE IN STRUCTURE OF POEM AND PROSE

With the age of information technology come large quantities of digitized text. Different techniques have been developed to categorize the available information. Word adjacency networks have been shown to reflect the difference in organization of scientific and literary texts (Grabska-Gradzinska et al. 13) as well as the distinction in the structure of poem and prose samples (Roxas and Tapang 503).

Prose, as the common form of written language, lacks the formal structure of meter or rhyme which is typical of poetry. Instead it is composed of full sentences and is usually divided into paragraphs. Figure 1 illustrates samples of word adjacency networks built from a poem and prose by Robert

Louis Stevenson. In the prose network, function words, such as articles, prepositions, and auxiliaries act as hubs. Punctuation marks in the text are the hubs in the poem networks. Network parameters that reflect the difference in the structure of poem and prose are the clustering coefficient, average path length and average degree.

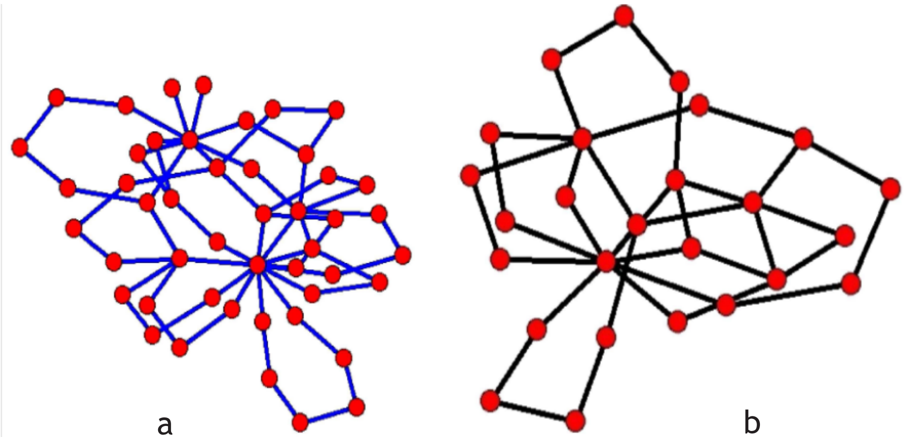

Figure 1. Word adjacency network of (a) the poem The Summer Sun Shone Round Me, with N = 80, C = 0.02, D = 2.69, and (b) the short story The Tadpole and the Frog, with N = 53, C = 0.18 and D = 2.93, by Robert Louis Stevenson. The short story has greater clustering coefficient and average degree even though it has lesser words indicating a more connected network compared to that of the poem. (Roxas and Tapang 506)

Using the same method, the change in the literary form within a single article can be automatically located. This was done by determining the clustering coefficient of the adjacency network of words inside a window that was shifted throughout the text. A sudden decrease of clustering coefficient indicates a switch from prose to poem while an abrupt change from low to high clustering coefficients signifies a transition from poem to prose. The text boundary falls within a window where the significant change in clustering coefficient is observed. This method can be useful for the text analysis of large databases of literary works.

Poetry sequence patterns and network motifs

The writing style and sequence patterns in English poems have been shown to change throughout literary history (Miles 7). Boundary lines are set to distinguish the change in writing styles through time. Using word cooccurence networks we probe if there exist a similarity within, and difference across, era as reflected within the poem's structure (Roxas-Villanueva, Nambatac and Tapang 1250009-6). English poetry was chosen as sample because it poses two advantages from a research perspective: well-represented and well-documented.

Three-node motif IDs (3-6, 3-12, 3-36), shown in Figure 2, were observed to have high frequency within the whole timeline. These motifs represent the most common subnetwork structure of the poems. Figure 3 shows the plot of the distribution of the poems with the motif ID frequency as the axis. Network motifs, which are considered the building blocks of complex networks, are patterns of interconnections that occur frequently. For each poem the frequency of each network motif was determined.

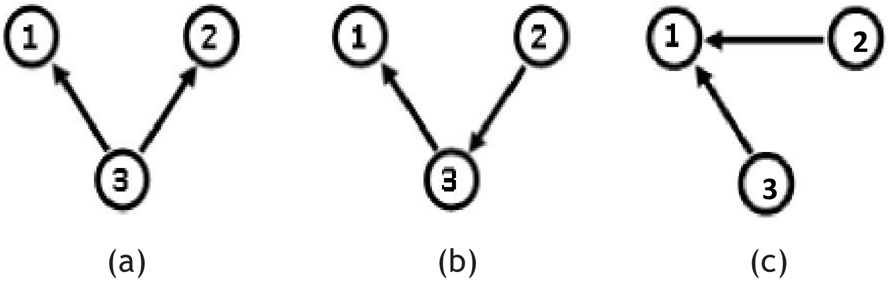

Figure 2. Most frequent three-node motif IDs, (a) 3-6, (b) 3-12, and (c) 3-36, in English poems published from 1522 to 1931. These motifs represent the most common subnetwork structure in the sampled poems. (Roxas-Villanueva, Nambatac and Tapang 1250009-5)

In Figure 3 the convex hull is the smallest polygon that can contain the data points (motif frequency of the poems) in each era. The similarity in structure of the Elizabethan, 17th Century, Romantic and Victorian eras can be observed in the close proximity of their centroid location and the overlap of their convex hull. The structural difference of Augustan poems is reflected by its smaller overlap with the rest.

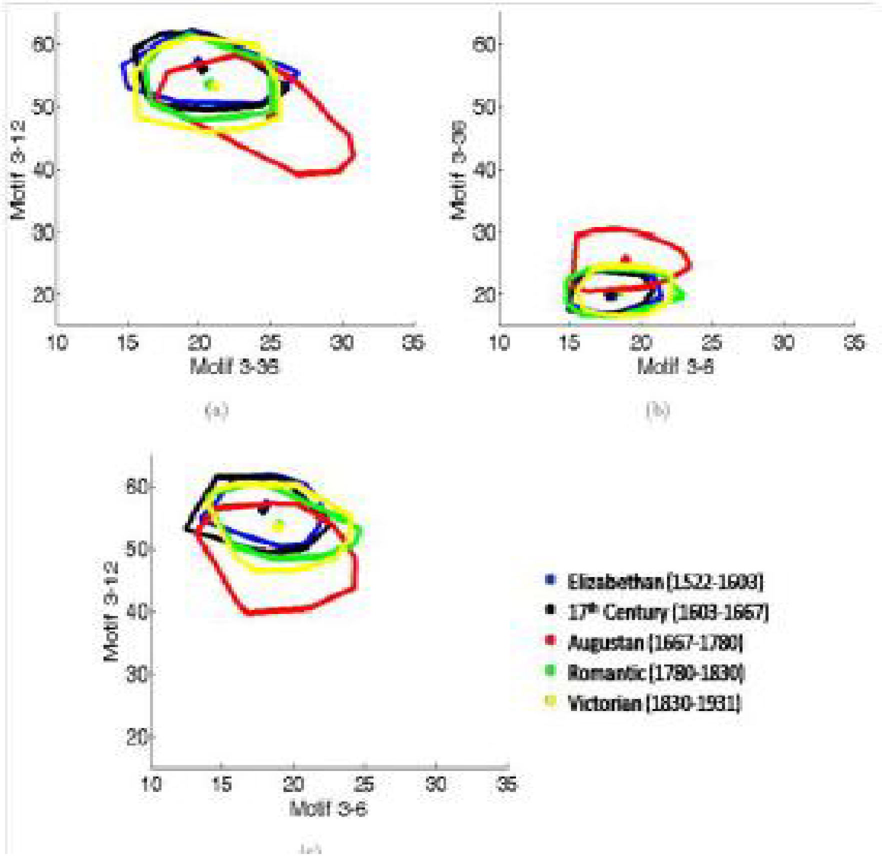

Figure 3. Convex hull and centroid of the distribution of the poems with the motif ID frequency as the axis for (a) 3-12 versus 3-36, (b) 3-36 versus 3-6 and (c) 3-12 versus 3-6 motifs. (Roxas-Villanueva, Nambatac and Tapang 1250009-6)

According to Miles there are three distinguishable sequence patterns into which the works of poetry from the 16th to 20th centuries can be categorized: adjectival, balanced and predicative (2). For example, ". . . an adjectival phrase would be formed as: rising and soaring, the golden bird flies into the stormy night of the east. While its clausal [predicative] version could be: the golden bird rises and soars; it flies into the night which storms into the east. . ." The balanced sequence pattern is a combination of both. Miles classifies poems of the Elizabethan, 17th Century, Romantic and Victorian eras as predicative and balanced while those written during the Augustan era as adjectival (Miles 7). The network patterns, properties and motif frequencies suggest the difference or similarity of the recurring sequential patterns in that particular era.

Network metrics of translated texts

Cabatbat et al built co-occurrence networks based on selected sets of chapters from different books of the Bible and the Universal Declaration of Human Rights (UDHR) in different languages (Cabatbat, Monsanto and Tapang 1350092-8). They compared these networks to random text networks. Among the network metrics considered, the network size, the normalized betweenness centrality and average k-nearest neighbor were found to be the most preserved across translations. Motif frequency distribution is also a viable tool for detecting text content similarity since motifs are found to be dissimilar in unrelated texts. Since translation is supposed to retain the meaning of texts across languages, these preserved network metrics establish a relationship between network structure and meaning. Metrics that do not change across different translations may be used as automatic classifiers or as means to quantify similarities in content of written texts.

More applications of complex network on literary analysis

By analyzing the metrics from complex networks representing texts from published books, Amancio et al investigated changes in the writing style over several centuries (Amancio, Oliveira Jr and Costa 043029). In this study, they treated books published from 1590 to 1922 as complex networks, whose metrics were analyzed with multivariate techniques to generate six clusters of books. The clusters correspond to time periods coinciding with relevant literary movements over the last 5 centuries. They observed that the most important factor distinctive of the different literary styles was the average shortest path length. Their results also showed there has been a trend toward larger average shortest path lengths, which is correlated with increased syntactic complexity, and a more uniform use of the words.

For literary works in the comic book medium, one study analyzed the Marvel Universe as a collaboration network where characters are connected if they appear in the same publication (Gleiser 10). Unlike most real social networks, the Marvel Universe is a disassortative network in which very dissimilar nodes (heroes and villains) are connected. Upon analyzing the connectivity of the nodes, they showed that the hubs are characters forming a group of heroes that connect different communities while characters labeled as villains appear around the hubs and do not connect communities. They accounted these to the rules of the Comic Code Authority which limit the role of villains. The connections between heroes mean teaming up which indicate that some effort is needed to defeat their enemies. The strongest link in the Marvel Universe is the relation between Spider-Man and Mary Jane Watson Parker illustrating that the most popular plot is a love story although the Marvel Universe is primarily about superheroes and villains.

Another related study analyzed the networks of three iconic mythological narratives: *Beowulf*, the *Iliad* and the *Táin Bó Cuailnge* (Mac Carron and Kenna 28002-p2). They define two distinct relationship types: friendly and hostile edges. To place the three mythological networks on the spectrum from the real to the fictitious, they compared their properties to actual and imaginary social networks. For comparison, they applied the network tools to four narratives from fictional literature: Hugo's Les Misérables, Shakespeare's Richard III, Tolkien's Fellowship of the Ring and Rowling's Harry Potter, which they found to have very high clustering coefficients, are disassortative and have giant components containing almost every character, indicating their societies' artificiality. Of the three myths, the network of characters in the Iliad has properties most similar to those of real social networks. It has a power-law degree distribution, is small world, assortative, vulnerable to targeted attack and is structurally balanced. The authors accounted this similarity to the archaeological evidence supporting the historicity of some of the events of the Iliad.

Archaeological evidence also suggests some of the characters in Beowulf are based on real people, although the events in the story often contain elements of fantasy. The network for this society, although consisting of a small number of nodes, has some properties similar to real social networks. It is, however, disassortative like all the fictional narratives; but it becomes assortative when the main character was removed from the network. The social network of the full narrative of the *Táin* initially seems similar to that of the Marvel Universe. However, its degree distribution is surprisingly similar to that of Beowulf, except for its top six vertices. This suggests that the network's artificiality may be mainly associated with the corresponding characters which are similar to the superheroes of the Marvel Universe —too super-human to be realistic, or in terms of the network, they are too well connected. The authors speculate that these characters may in fact be based on amalgams of a number of entities and proxies. To test this hypothesis, they removed the weak social links associated with the six characters and this resulted to a network that is assortative, similar to the Iliad and to other real social networks and very different to that of the Marvel Universe and works of fiction.

The recent increase in network studies research can be linked to the considerable increase in computing power and network infrastructures and to the heightened access to enormous data facilitated by computer and Internet technologies. However, the study of complex networks and its applications is still in its development stage. Much can still be analyzed using such technique. Possible further applications are authorship attribution of disputed literary pieces, poetry form and rhyme scheme analysis, and the analysis of other art forms like music.

Works Cited

Alvarez-Lacalle, Enrique, Dorow, Beate, Eckmann, Jean-Pierre, and Moses, Elisha. "Hierarchical structures induce long-range dynamical correlations in written texts" *P. Natl. Acad. Sci. USA*, 103(21) (2006): 7956-7961.

Amancio, Diego Raphael, Oliveira, Osvaldo Jr., and Costa, Luciano. "Identification of literary movements using complex networks to represent texts." *New J. Phys.* 14 (2012) 043029.

Barabási, Albert and Albert, Réka. "Emergence of scaling in random networks." *Science*. 286 (1999): 509-512.

Cabatbat, Josephine Jill and Tapang, Giovanni. "Texting Styles and Information Change of SMS Text Messages in Filipino." *Int J Mod Phys C*. 24 (02) (2013).

Cabatbat, Josephine Jill, Monsanto, Jica, and Tapang, Giovanni. "Preserved network metrics across translated texts." *Int. J. Mod. Phys C* **25(2), (**2014**).**

David, Clarissa, Atun, Jenna Mae, Legara, Erika Fille and Monterola, Christopher. "Finding Frames: Comparing Two Methods of Frame Analysis." *Communication Methods and Measures.* 5(4), (2011): 329-351.

Ferrer i Cancho, Ramon, and Solé, Ricard. "The small world of human language." *Proceedings of the Royal Society of London Series B, Biological Sciences*. 268(1482) (2001):2261-2265.

Ferret, Olivier. "How to Thematically Segment Texts by using Lexical Cohesion?" *Proc. 36th Annual Meeting of the Association for Computational Linguistics*. (1998): 1481.

Gao, Yuyang, Liang, Wei, Shi, Yuming, Huang, Qiuling. "Comparison of directed and weighted co-occurrence networks of six languages." *Physica A*. 393 (1) (2014): 579–589.

Gleiser, Pablo. "How to become a superhero." *J. Stat. Mech.* (2007) P09020.

Grabska-Gradzinska, Iwona, Kulig, Andrzej, Kwapien, Jaroslaw, and Drozdz, Stanislaw. "Complex network analysis of literary and scientific texts." *Int. J. Mod. Phys. C* 23(7), (2012).

Kapustin, Victor and Jamsen, Anna. "Vertex degree distribution for the graph of word co-occurrences in Russian." In *Proceedings of the Second Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing*, pages 89-92, Rochester, NY, USA, 2007. Association for Computational Linguistics.

Legara, Erika Fille, Monterola, Christopher, and David, Clarissa. "Complex Network Tools in Building Expert Systems that Perform Framing Analysis." *Expert Systems With Applications.* 40(11), (2013):4600–4608.

Legara, Erika Fille, Monterola, Christopher, David, Clarissa, and Atun, Jenna Mae. "News framing of population and family planning issues via syntactic network analysis." *Intl J Mod Phys C* 21(1), (2010):51-65.

Liang, Wei, Shi, Yuming, Tse, Chi, Liu, Jing, Wang, Yanli, and Cui, Xunqiang. "Comparison of co-occurrence networks of the Chinese and English languages." *Physica A*. 388 (23) (2009): 4901–4909.

Miles, Josephine. *Eras and Modes in English Poetry*. Berkeley: University of California Press, 1964.

Motter, Adilson, de Moura, Alessandro, Lai, Ying-Cheng and Dasgupta, Partha. "Topology of the conceptual network of language." *Physical Review E*. 65 (2002.) 065102:1-4.

Pádraig Mac Carron and Ralph Kenna. "Universal properties of mythological networks." *EPL* 99, (2012) 28002.

Pastor-Satorras, Romualdo and Vespignani, Alessandro. "Immunization of complex networks." *Phys Rev E*. 65 (2001): 036104.

Roxas, Ranzivelle Marianne and Tapang, Giovanni. "Prose and poetry classification and boundary detection using word adjacency network analysis." *Int. J. Mod. Phys. C* 21(4), (2010).

Roxas-Villanueva, Ranzivelle Marianne, Nambatac, Maelori Krista and Tapang, Giovanni. "Characterizing English poetic style using complex networks." *Int. J. Mod. Phys. C* 23(2), (2012).

Steyvers, Mark, and Tenenbaum, Joshua. "The large-scale structure of semantic networks: statistical analyses and a model of semantic growth." *Cognitive Science*. 29(1) (2005): 41-78.

Watts, Duncan and Strogatz, Steven. "Collective dynamics of small-world networks." *Nature*. 393 (1998): 440-442.