Text Network Analysis

by Dmitry Paranyushkin, April 2010

This article was adapted from a talk given at Performing Arts Forum (PAF) in France in April 2010.

In this article I would like to propose a method for network text analysis. This approach is different from semantic text analysis, which focuses on interpreting the relations between the terms based on their meaning [1].

If we represent text as a network of words and key concepts that are related, then semantic analysis would add an extra layer of ontologies describing the relations and, thus, increasing the complexity (e.g. "is this a negation?", "are these two people friends?", etc). Such approach has many advantages, but also subordinates the resulting interpretations to the logic of language.

Instead, the method proposed here focuses on the topology of networks and their structure in order to gain interesting insights about the text. This more visual approach has several advantages.

First, we can see a text at once, as a Gestalt, instead of waiting for it to unfold in front of us with time. "Time prevents everything from being given at once." [2]. Visualizing text as a network we remove the variable of time and let the history of the text appear through the diagram.

Second, such visualizations help us trace the pathways for meaning circulation within the text. Obvious and latent power centers within the text as well as hidden agendas can easily be detected this way.

Third, network analysis goes way beyond traditional text analysis tools. We are not interested in tag clouds or finding the most frequently mentioned concepts. Instead, we can analyze the actual relations, the processes that align the words in a specific way, rather than the terms themselves.

Finally, network analysis can unlock the potentiality of a narrative. Italo Calvino once said that writing is essentially a combinatorial exercise and reading is "a way of exercising the potentialities contained in the system of signs" [3]. Representing text as a network does exactly this: offers many possibilities for interpretation and reading, which would normally be suppressed by the dominant narrative.

In order to demonstrate how text network analysis works, I will use the actual text above and make a network representation of it.

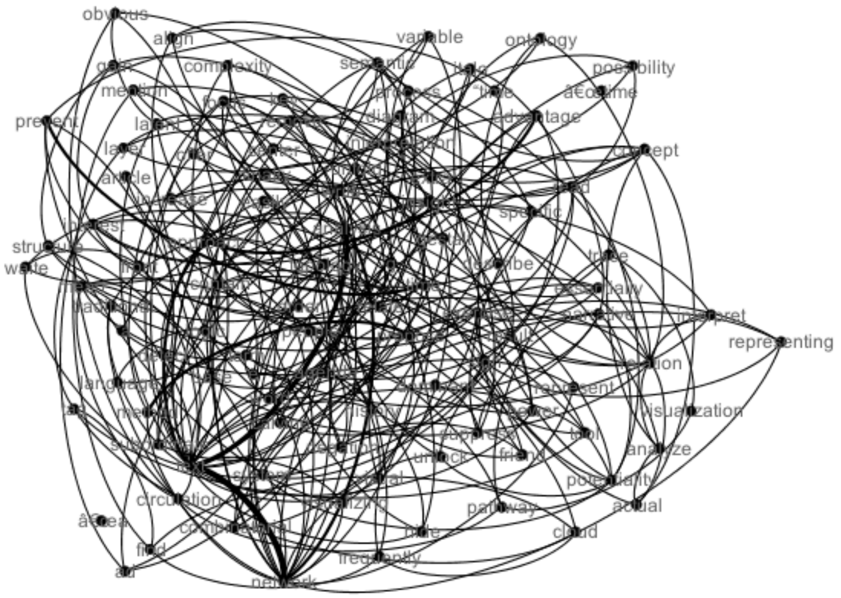
Using AutoMap software [4] the text was simplified by removing all of the stopwords (a, the, is, are, etc.) and applying stemming, which normalizes the past and the future tense to the present (so that words like "writing", "wrote", and "written" would appear in our network as the word "write"). This helps decrease the complexity and reduce the amount of processing required. The resulting text looks something like this:

> article propose method network text analysis approach semantic text analysis focus interpret relation term base mean represent text network word key concept relate semantic analysis ad extra layer ontology describe relation increase complexity e g negation people friend approach advantage subordinate result interpretation logic language method propose focus topology network structure order gain interest insight text visual approach advantage text gestalt waite unfold front time time prevent visualizing text network remove variable

> time history text diagram visualization trace pathway mean circulation text obvious latent power center text hide agenda easily detect network analysis traditional text analysis tool interest tag cloud find frequently mention concept analyze actual relation process align word specific term finally network analysis unlock potentiality narrative italo calvino write essentially combinatorial exercise read a exercise potentiality contain system sign representing text network offer possibility interpretation read suppress dominant narrative

Next, a csv file is generated listing the keywords in context, as well as a semantic network in xml format showing the relationships between the words. In order to generate a semantic map we take "windows" of text 3 words long and find how often they appear together. The resulting files are converted using ThisIsLike software [5] and present a map of word co-occurences, which allow us to build a network representation of the text.

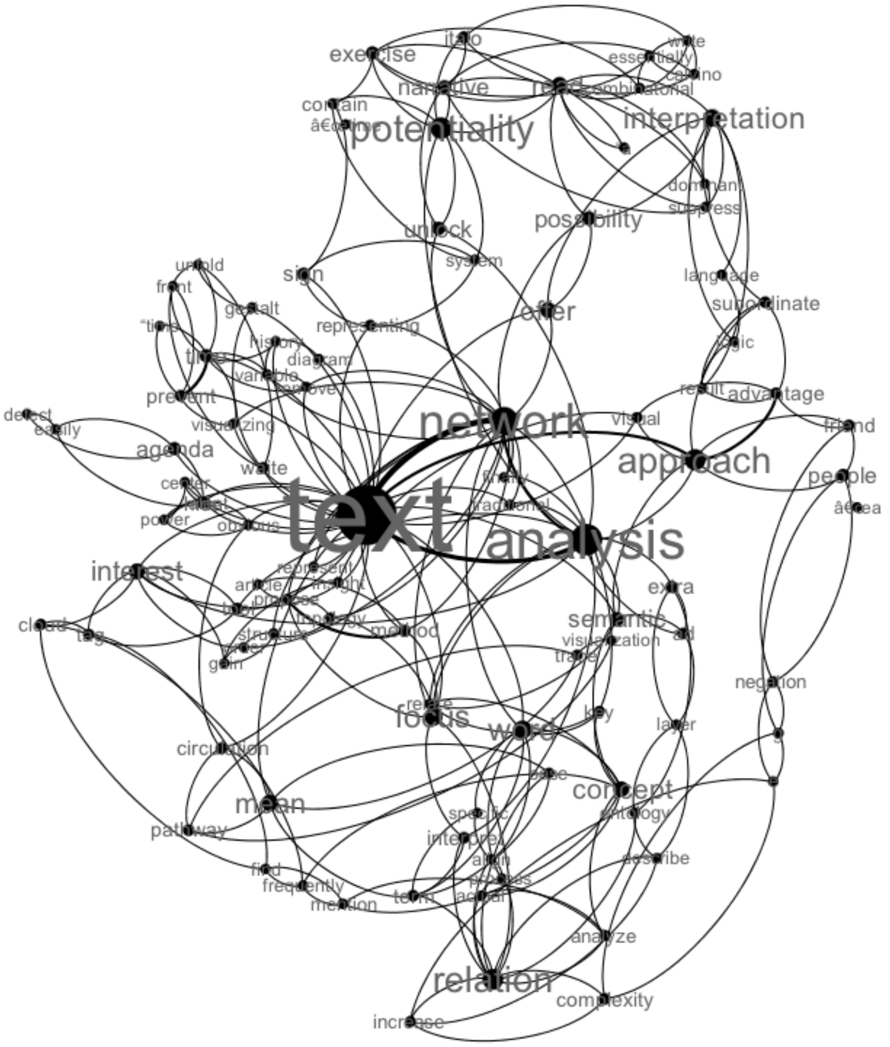
The both files are then imported in Gephi network visualization software [6] and the image of the network is generated.



At first, the nodes are aligned in a random way. The goal is to obtain meaningful information from this image, so we range the nodes by their betweenness centrality (or influence) and detect the communities.

A video of this text network visualization in Gephi can be seen on http://vimeo.com/16692084

One of the most interesting properties of a node is its "betweenness centrality". It is a measure of how often a node (or a word, in our case) appears on the shortest path between two other random nodes in the network. In other words, this measure shows the most influential nodes that connect different communities within the network. The higher it is, the more often the word connects various contexts within the text. These are shown bigger on the graph below.



The words with the highest betweenness centrality are:

text, analysis, network, potentiality, approach, relation

Not all the nodes which have many connections (high degree) to other nodes can be considered "influential". For example, a node may have many connections by forming a cluster of nodes around itself, and yet not being fully integrated into the larger network.

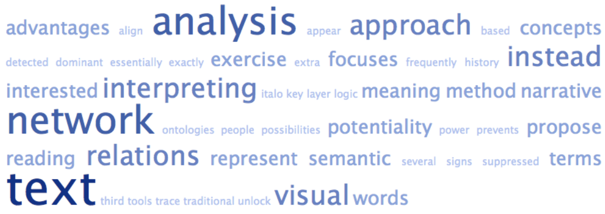
In other words, betweenness centrality shows the variety of contexts where the word appears, while high degree shows the variety of words next to which the word appears. Texts with the highest degree (or the number of connections to other words) are:

text, network, analysis, relation, time

It can be seen that the word "time" is mentioned next to many other words in the text, however, it is not a very influential word, because it's only mentioned in a specific context and not throughout the whole text itself.

On the other side, "potentiality" is only mentioned once in the whole text, but it's a very important word, because it connects different contexts together (network analysis and unlocking potentiality of a narrative through reading).

It's important to point out the difference of network text analysis from keyword frequency analysis and tag clouds. First, keyword frequency does not allow us to detect clusters of meaning within the text, neither visually, nor quantitatively. Second, tag clouds do not tell us anything about the relations between the words, rather, the relation of a specific word to the whole text only. A tag cloud for the same text would look something like this [6]:

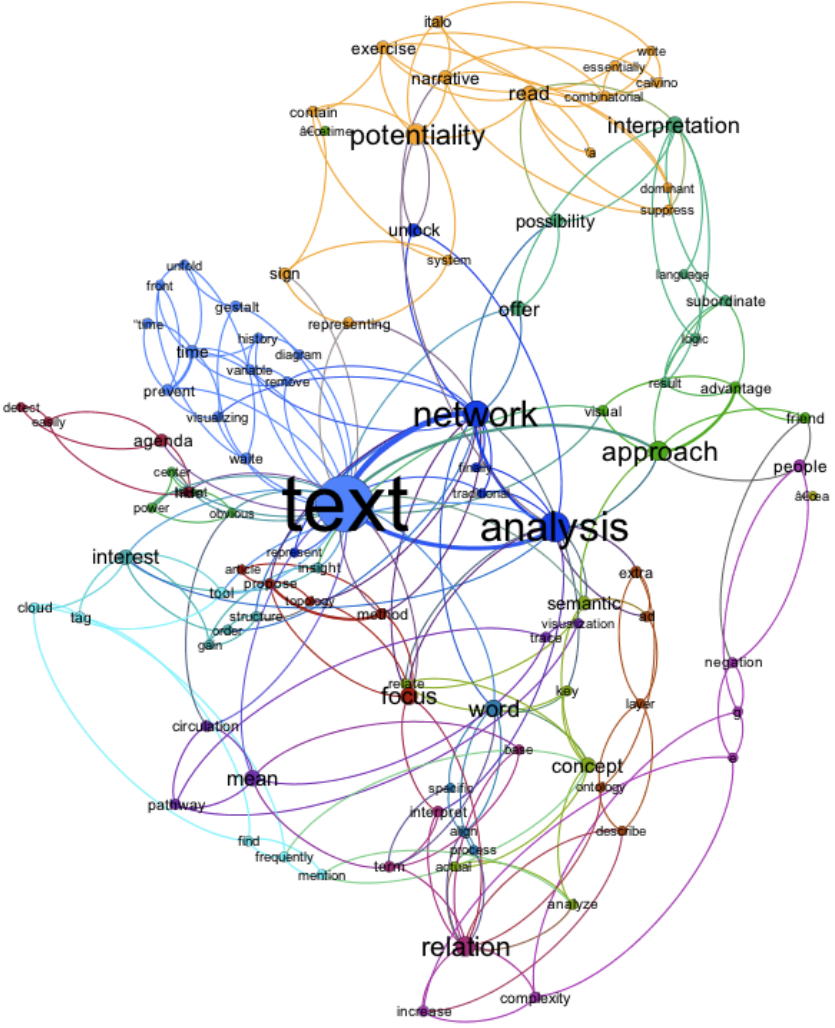


We can still see the main words, but don't at all see any communities, let alone relations between them.

Text network visualization can also give us insights into the dominating agendas within the text. Looking at the network, we can detect the main loop of meaning circulation:

text - network - analysis - text

We can also detect several communities or clusters of meaning – the nodes (or words, in this case) that are better embedded into their neighborhood [7]. The algorithm utilized in Gephi, which follows a method proposed by Blondel et al [8], detects two major communities within the network:



The main community includes about 16% of the words in the text, among which the most "influential" are:

potentiality, read, exercise, narrative

The other community includes about 13% of the words in the text, among which are:

text, time, prevent, diagram

Therefore, we can now identify dispositif of the text as an interaction between two clusters of meaning: one preoccupied with potentiality of a reading and narrative, the other – with preventing temporal unfolding of the text and presenting it as a diagram instead. This interaction happens through the loop "text - network - analysis", which is the proposed method to relate the two communities together.

The other communities on the periphery serve to exemplify the method. For instance, the next largest community includes the words "interpretation" and "possibility" and acts as a pathway between the dominant clusters.

The strategy here, therefore, is to

1) identify the loop(s) and pathways for meaning circulation (words with highest betweenness centrality that are connected together) – this can also be called agenda;

2) identify the communities or clusters of meaning within the text;

3) explore the relations between the communities of meaning;

4) find alternative pathways through which these communities connect bypassing the loop for meaning circulation – these will usually exemplify the main agenda, but in different terms;

In other words, the proposition of this article judging from the text network analysis above is to unlock the potentialities in reading through removing the time as the determinant factor for the way we normally perceive a text. This can be done through network analysis, which implies playing with various possibilities for interpretation that this text offers.

But there could be many other readings, of course. One could follow a totally different pathway and find a totally different interpretation. This distance between the possible meaning and various ways of making sense is exemplified in the image of a network. The diagram acts as a dysfunctional interface, which compresses the time into one moment, opening it up for multiple interpretations. As much as it is arbitrary, it is also relieving in the sense that it shows that all (mis)communication is a matter of punctuation and sequencing as much as it is a matter of the actual information transmitted.

"The nature of a relationship is contingent upon the punctuation of the communicational sequences between the communicants." [9]
When we read a text, we follow a narrative guided by the author, rules of grammar, logic, and common sense. When we read a network, we follow the affirmative drive of contingent associative flow.

Resources:

[1] Wouter Van Atteveldt, "Semantic Network Analysis: Techniques for Extracting, Representing, and Querying Media Content" (2008, Booksurge LLC)

[2] Henri Bergson, "The Possible and the Real" (2002, Continuum)

[3] Italo Calvino, "The Uses of Literature" (Orlando: A Harvest Book, 1986)

[4] Automap by Kathleen M. Carley at CASOS (Center for Computational Analysis of Social and Organizational Systems) at Carnegie Mellon University - www.casos.cs.cmu.edu/projects/automap/software.html

[5] ThisIsLike - www.thisislike.com/utils/xml2graphml.php

[6] TagCrowd - www.tagcrowd.com

[7] Santo Fortunato, "Community Detection in Graphs" (2010, Complex Networks and Systems Lagrange Laboratory, Torino)

[8] Blondel et al, "Fast Unfolding of Communities in Large Networks" (2008)

[9] Watzlawick et al, "Pragmatics of Human Communication" (1967, W W Norton & Company)