How to Explore a Billion Social Media Entries and Tell a Story!

Neo Mohsenvand
MIT Media Lab, Social Machines
Research Proposal

What are the goals and their importance?
This project-attempts to explore and summarize large amounts of Twitter, Facebook and other social media platform posts and replies in the following ways:

1. Provide a visual analytic approach for semantic navigation through a large set of data entries using an obtained or induced ontology tree;
2. Design and evaluate a story telling or report generation scheme (at bullet point level) based on the hierarchical community structure of the data

The inspiration for this project emerges from the ways in which humans intuitively understand big heterogeneous data through ontologies and hierarchies [1]. The author hypothesizes that knowledge and meaning are defined in the context of networks and can be understood and internalized using hierarchical community structures (Appendix A).-These structures are here represented by community metagraphs and ontology trees. In one case, I demonstrate that multiple ontologies of a data set can simplify the semantic search process (Appendix B). In another, I develop a navigation tool for visualizing trends using the ontology tree, also enabling users to compare ongoing trends when the number of time series is very large (Appendix C).

As for story telling, a descriptive narrative of a data set can be conceptualized as a traverse through the extracted ontology tree (appendices D and E). One observes that, similar to the trend visualization tool, parent nodes of the tree have access to aggregated or summarized information coming from the child nodes. Such traversals can remarkably provide a continuous narrative in a bulleted fashion. These bullet-points will then be stitched together using an NLG algorithm to generate a human readable text. This ontology traversal scheme also allows for change in length, sentiment and emphasis of the report using the control parameters of the traversal algorithm.

How can we address these goals?
General Steps

1. Extracting networks from posts and replies and related attributes and annotations where nodes and links are enhanced with different properties. This part of the project is going to be implemented as a graph database (using Neo4j and Cypher) so that other lab members can use it for their own queries;
2. Finding community structure by efficient implementation of various algorithms

Goal 1 specific steps:

1. Build the underlying software infrastructure for visualization of trees based on the graph database and community detection analysis;
2. Come up with a comprehensive set of useful aggregation and curation methods to enable various modes of visualization

Goal 2 specific steps:

1. Design various traversal algorithms with tunable parameters to generate structured summaries;
2. Explore NLG algorithms to turn the bullet-points sequence into a narrative

What data and other resources are needed?
A large amount of data entries with meta- data and possibly the underlying user network

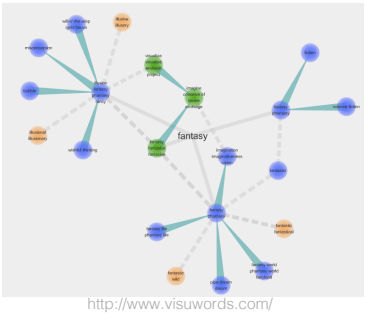
How will I measure success?
Outputs of the project are two products that will be used by humans. Therefore, besides the software engineering cycle, one has to specify some usability and accuracy criteria to evaluate the success of the project. Many such evaluations have already been developed and studied in visual analytics and HCI communities. Amongst these, the Creativity Support Index (CSI) and the NASA Task Load Index (TLX) can be seen as useful measures [2,3].

Appendix A : Philosophy

Meaning is a Network Property

Raw data is like a corpse that is not able to communicate. On the other hand, a network is a living and responsive entity.

Thinking deep, we always define complex ideas in relationship to other things. The word “school,” for example, can only be defined through the network of related concepts and meanings such as “teacher,” “student,” “lesson,” “exam,” etc. We can not even define properties of physical objects such as velocity, position, even mass, spin and electric charge of elementary particles without considering their interactions with other particles. For a single free electron in a completely empty space it is impossible to define any physical properties or even argue about its existance. This idea is called Mach’s principle, and has influenced me in understanding the nature of meaning for general kinds of data.

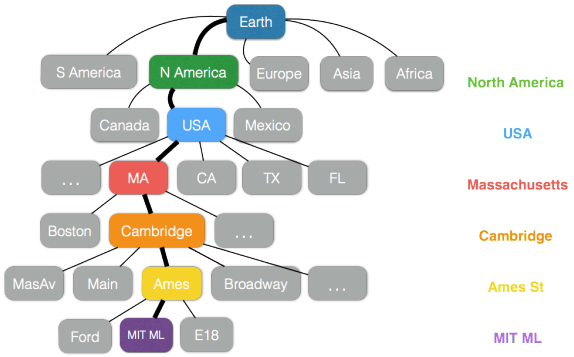


http://www.visuwords.com/


Ernst Mach , 1916

Humans Understand Complex Things Through Hierarchies

If you try to write down the names of the cities you heard in your life, you can list pages of city names. In contrast, remembering a dozen phone numbers or a list of cities of a fictional planet is a much more challenging task.
However, if instead of a mere list, we are given a hierarchical structure or a tree of names, the first level being a list of continents, the second level a list of countries in each continent and the third level a list of cities in each country, something magical happens.



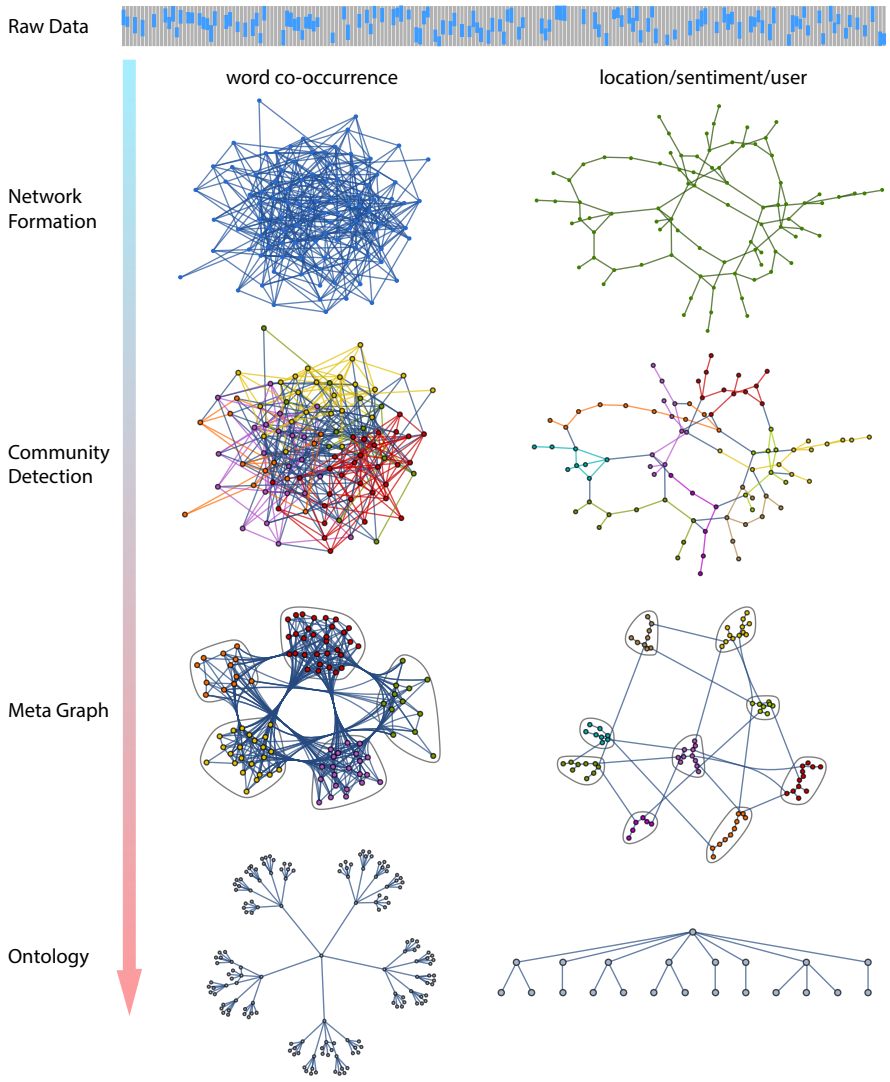
We humans use hierarchies to understand complex things. We assign symbols to different scales and learn the tree of symbols. Many believe that this is due to the hierarchical structure of the neocortex. I used this idea as the foundation for a tool that takes the extracted networks of data and finds communities in the network in a recursive manner.

Hierarchical Aggregation and Visualization

By augmenting raw data with associated networks and hierarchies, one can provide an interface or language with which to semantically communicate and explore.
Extracting the network of phrase co-occurrence provides a hierarchical topic model for the data set. In another example, a network of citations (such as direct and indirect retweets) can help us categorize different eras of an event over time (see Appendix E).

Community detection (CD) in networks provides a mesoscopic (therefore system-level) view for understanding their organization and function. If applied recursively, CD will provide us with a tree structure that encompasses the inherent hierarchy of assemblies in a network. Such hierarchy can help us navigate through different levels of aggregation and abstraction.

The figure on the right illustrates the process of generating meta graphs and ontologies from raw data. Here we are taking a set of faceboo post or tweets as the raw data and form two different networks based on the content (word co-occurrence) or attributes associated with the posts (say sentiment, geotags, user network etc).
It should be noted here that every node in these networks correspond to a single or multiple posts.



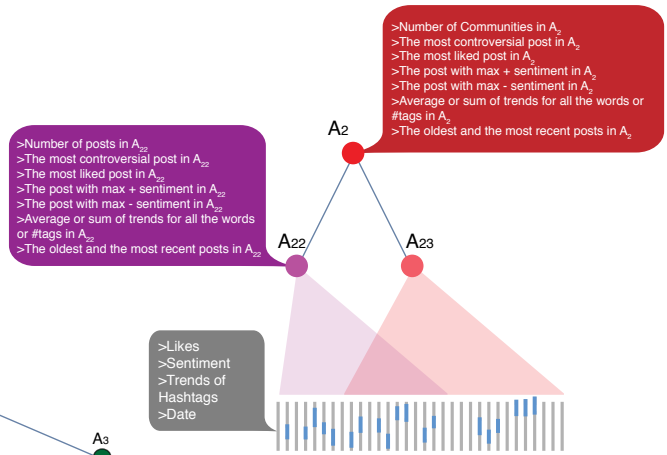
Neo (Mostafa) Mohsenvand, MIT Media Lab

Appendix B : Semantic Search Using Hierarchies

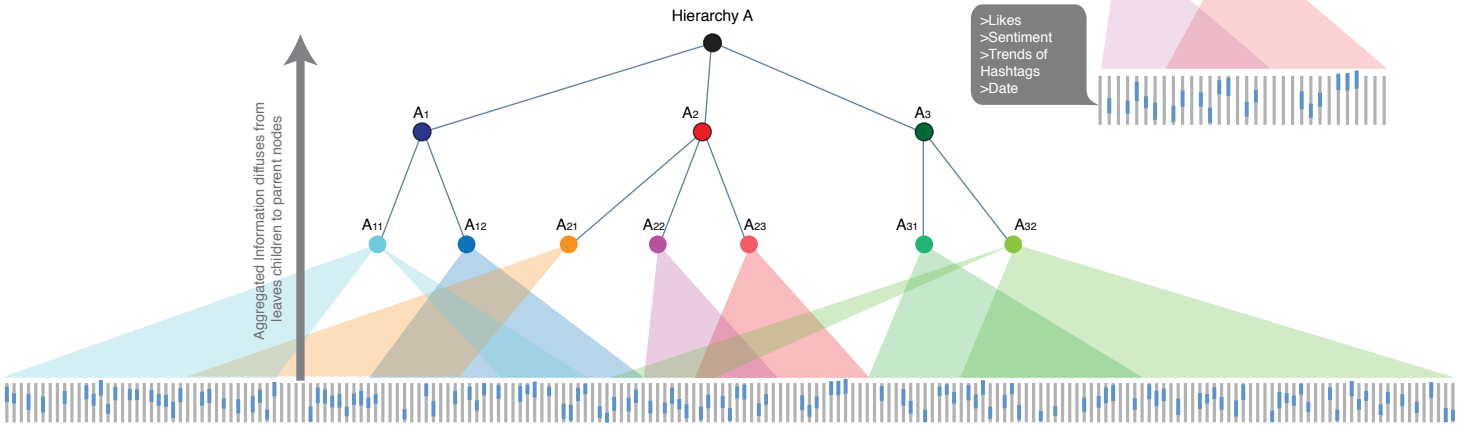
Here we show how an ontology enables us to semantically navigate through a dataset and summarize/aggregate the information at various levels of abstraction. This is made possible by using the concept of **Bottom-Up Information Diffusion (BUID)** which simply means that parent nodes contain a summary or aggregation of all the information in their children.

This simple rule demands a recursive aggregation of information all the way up to the root node. As a consequence, the root node has a summary of the whole data set.

The figure on the right shows an example of BUID where the aggregated information and stats of individual posts has diffused to the sub-communities and communities.



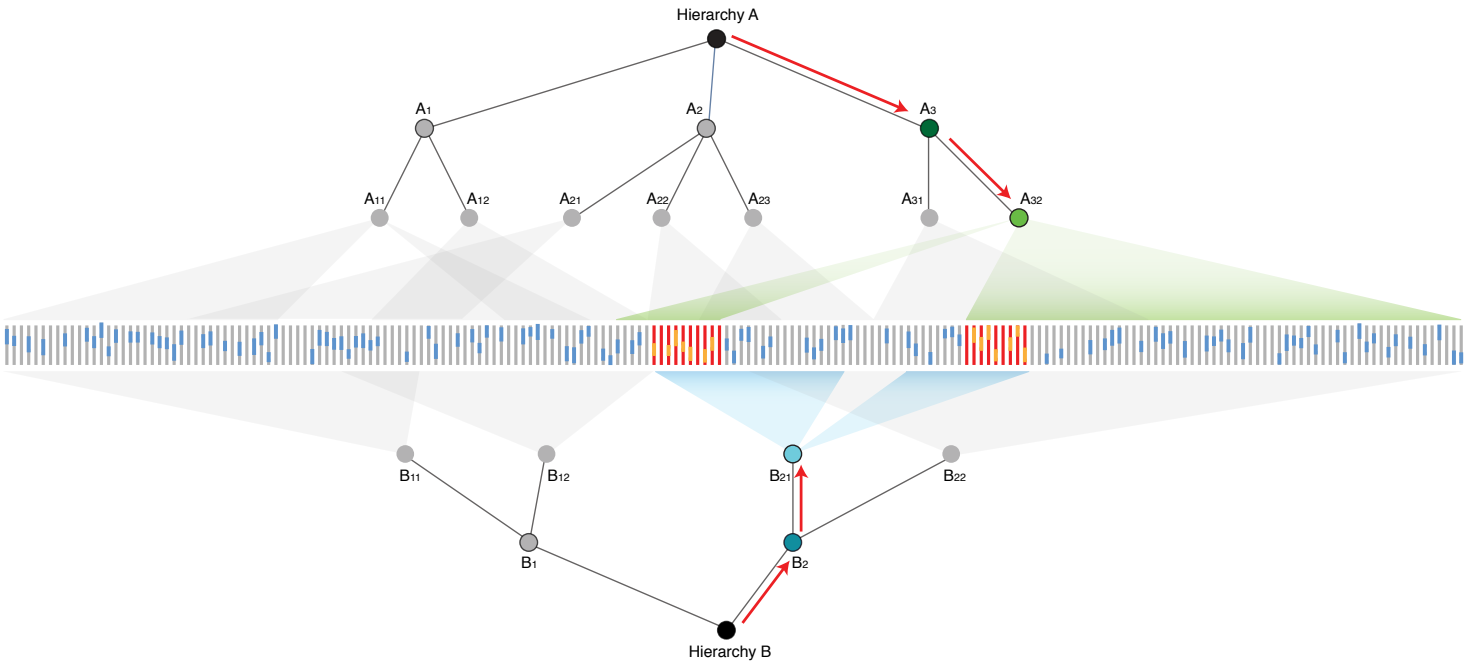
The figure below show an entire hierarchy on a set of posts.



BUID enables us to assign meaningful labels or properties to the intermediate nodes. In Appendix C we demonstrate an example of such labeling. One can use these labels to semantically navigate through the network. For example, take the ontology tree for the hashtag co-occurrence network extracted from the data set of tweets on the Delhi Rape Case. If one is only interested in the tweets that represent the opinion of foreign news networks on the issue, by looking at the ontology tree, one will find a category of hashtags labeled as News, Pakistan, USA. Choosing this subcommunity narrows down the search space and gives a semantically coherent set of tweets.

One can use multiple networks to further specialize the search from two different viewpoints. For example, the second network may have come from the community detection on the followership network of users. Then one can search for all the tweets about news by, say, conservative British journalists on the Delhi rape case.

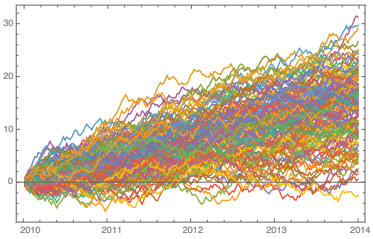
The figure below shows how using two different ontologies can further narrow down the number of resulting posts from the semantic search. The red highlighted posts show the intersection of two queries on hierarchies A and B.



Appendix C : A Visual Analytics Tool

Imagine we are interested in comparing the popularity of a set of hashtags over time in a specific set of tweets. Fig. A shows the trends for say 200 hashtags (generated by a brownian motion). It is evidently difficult making a sense of trends when the number of time series is large. However, by using the hashtag-co-occurrence network (Fig. B) which is extracted from the tweets, one can construct an ontology (Fig.C) for the hashtags . ***Here the community labels are determined by the labels of children in that community that have the highest page rank (an example of Bottom-Up Information Diffusion).***

Fig. A



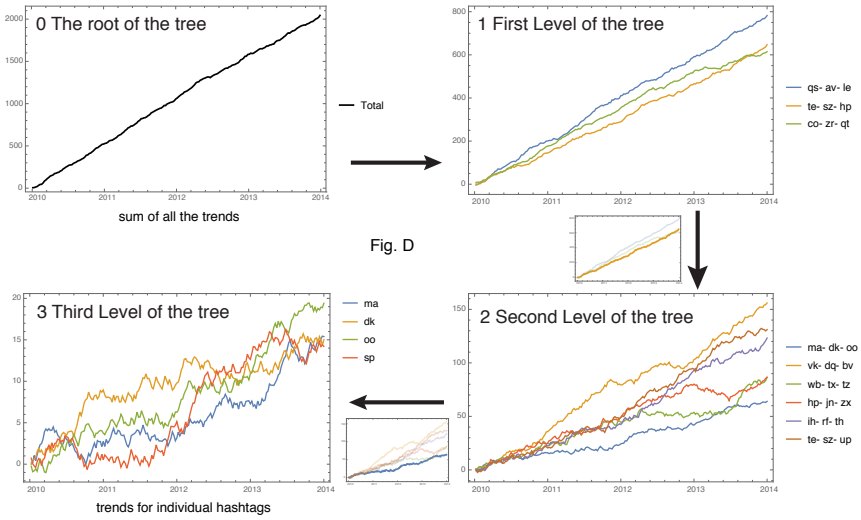
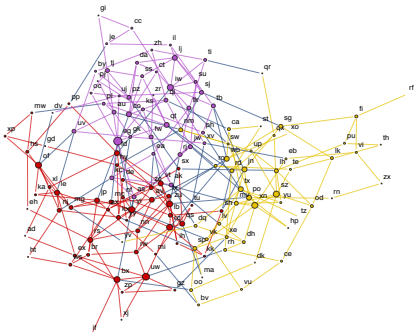


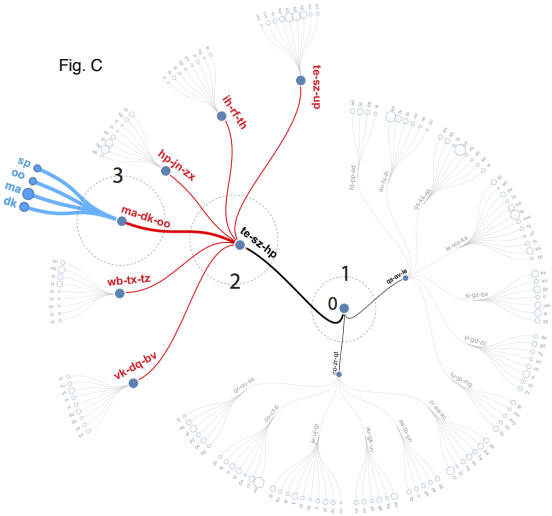
Fig. D

Fig. B



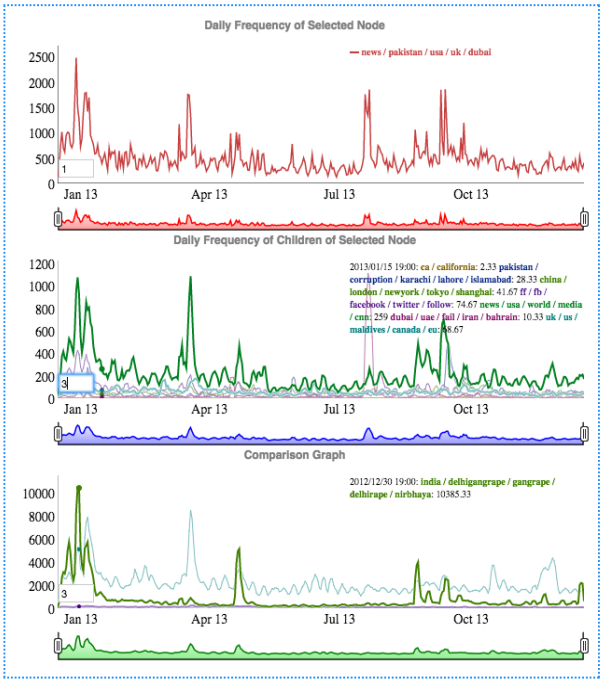
Here the size of the nodes shows their pagerank. Label of each community is composed of the labels of the nodes with top 3 pagerank.

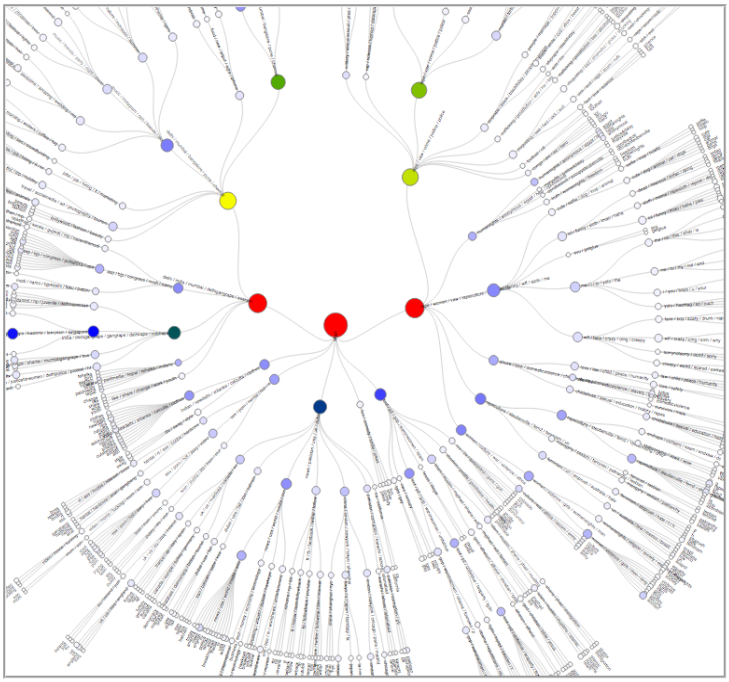
Fig. C



Now by simply taking the trends of the hashtags as an attribute for each tag and applying BUID by recursively aggregating them over the communities we will enable the user to obtain an overal image of the trends by navigating through the tree(Fig.C). In particular, when the user expands a node on the tree , the aggregated trends for the children of that node will be shown. Fig.D demonstrates the resulting trends from the four stages of navigation marked on Fig.C.

Based on this simple idea, I built a tweet explorer app that could aggregate the trends of tweets based on their assigned topic (obtained by community detection). The Screen-Shot below shows the tree obtained by analysis of 50 Billion tweets from a year-long activity related to the indian rape case of 2012.





Neo (Mostafa) Mohsenvand, MIT Media Lab
Facebook Fellowship Appendix C

Appendix D : Ontology Traversal as an Organic Storyline

The idea: Traversing the ontology tree in conjunction with the meta graph provides an organically styled storyline about the dataset from which the network is extracted.

In order to keep the continuity of the story, two generic moves are allowed: 1- Moving along the edges of the tree (=>), and, 2- level-order jumps (-->J -->) (see Fig 2). Level-order jumps are not allowed however unless there exists at least one corresponding link between the parallel communities on the meta graph.

Every passed node or jump produces a set of bullet points depending on the information on the node or the content of the intercommunal link on the meta graph. List 1 is produced by a particular traversal of the ontology tree whereas List 2 shows the traversal driven (parametrized) by three important examples in the data (t_1,t_2,t_3), the allowed length of the tour and the expected structure (say beginning and ending at the main nodes). This parametrization may be changed in order to produce other kinds of narratives (see Apendix E).

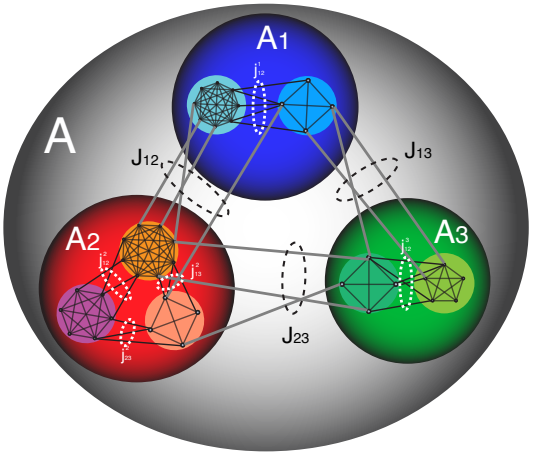


Fig 1

The main node (A) has access to the aggregated information of all the tree and is therefore a natural starting point. However, changing the parameters can result in a different initial point (see Story 4 below).

List 1

>Introduction
>aspect 1
supporting detail 1
link 1
supporting detail 2
example 1
>link 2
>aspect 2
supporting detail 3
link 3
supporting detail 4
example 2
link 4
supporting detail 5
>link 5
>aspect 3
supporting detail 6
link 6
supporting detail 7
example 3
>conclusion

List 2

$A \Rightarrow A_1 \Rightarrow A_{11} \rightarrow j^1_{12} \rightarrow A_{12}$

$\Rightarrow (t_1 \in A_{12}) \rightarrow J_{12} \rightarrow A_2 \Rightarrow A_{21} \rightarrow j^2_{12} \rightarrow$

$\Rightarrow A_{22} \Rightarrow (t_2 \in A_{22}) \rightarrow j^2_{23} \rightarrow A_{23} \rightarrow J_{23} \rightarrow A_3$

$\Rightarrow A_{31} \rightarrow j^3_{21} \rightarrow A_{32} \Rightarrow (t_3 \in A_{32}) \Rightarrow A$

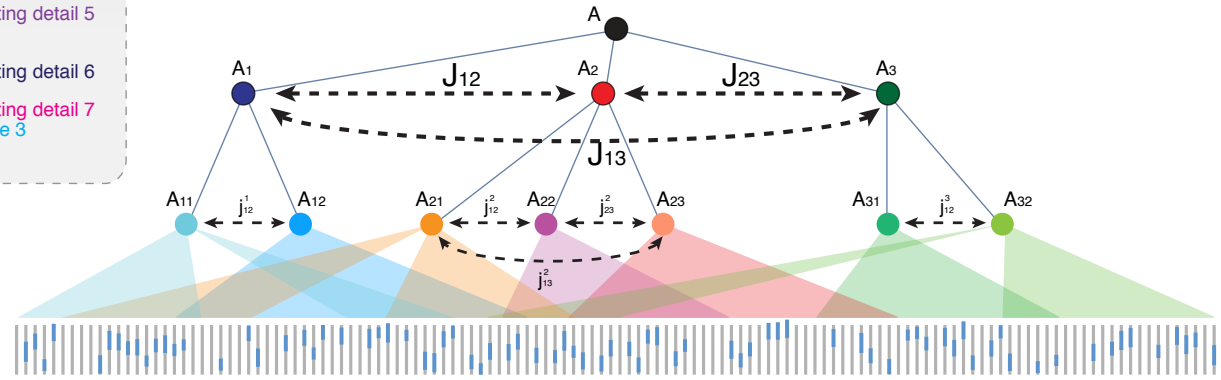
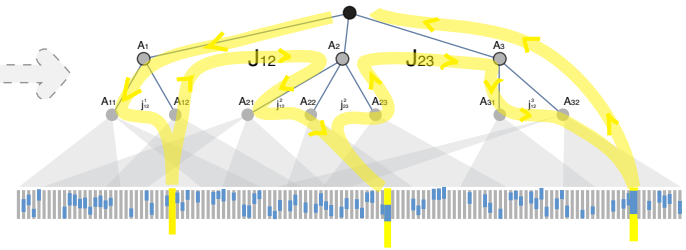
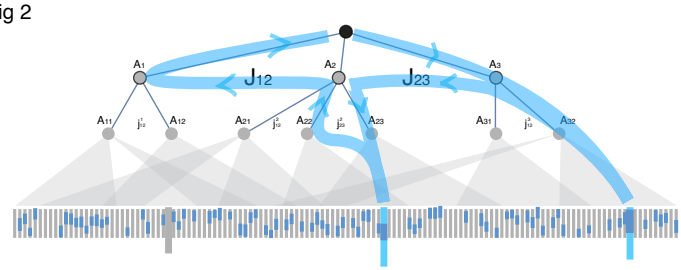


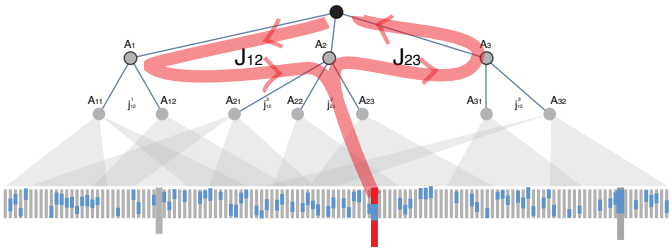
Fig 2



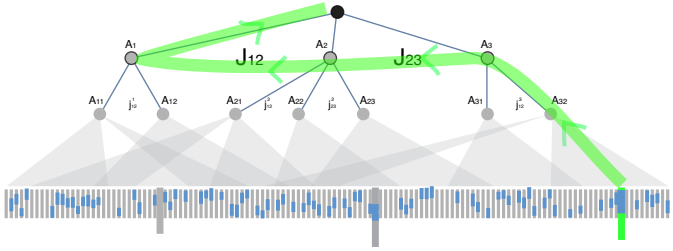
Story 1: This is based on the traversal shown in List 2. Constrain on the length is almost relaxed and a comprehensive narrative is generated.



Story 2: In this case, the length constraint has overcome the sensitivity to the t1 example (say it has a lower PR compared to t2 and t3); thus, a shorter story is generated.



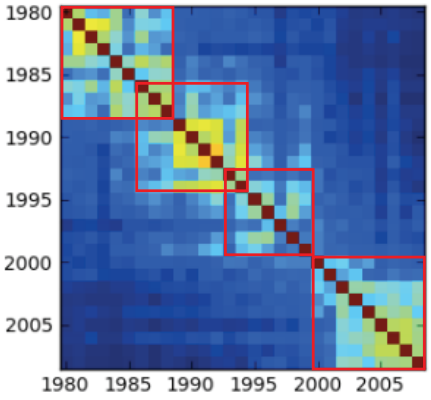
Story 3: Here the length constraint was much tighter, resulting into a single layer overview with a mention of the most important example.



Story 4: This parameterization is completely different from the previous ones due to it not enforcing the essay structure. Instead, the story starts from a specific example moves towards more general information.

Neo (Mostafa) Mohsenvand, MIT Media Lab
Facebook Fellowship Appendix D

Appendix E : Inducing Temporal Structure on the Narrative



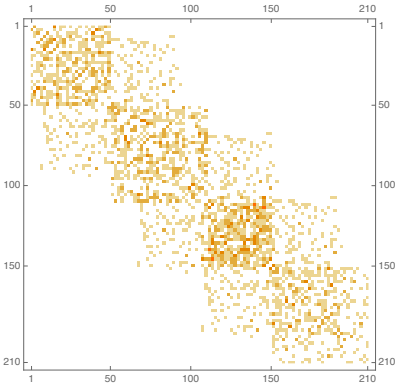
Narratives with a monotone direction of time-- such as sport/event reports, biographies and histories-- are so common. In this section, we show that the ontology based method can be enhanced to capture the direction of time. Although using the timestamp as a parameter enforces a temporal order to the traversal path, it does not change the community structure accordingly.

It turns out that for some datasets (especially those gathered over a long time), the data itself expresses a temporal community structure. In these cases, the obtained communities are representative of different eras of the evolution of the dataset.

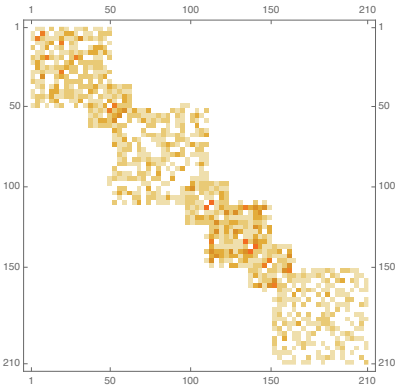
The figure on the left which is taken from (Jurafski et. al. ACL-2012) shows a temporal community structure in the topics discussed in NLP papers between 1980-2008. It shows how four almost distinct eras can be detected by a community detection algorithm.

Here we demonstrate a basic trick in order to accentuate the temporal community structure. This can help one get a monotone narrative over time that abstracts and aggregates over major eras in the process.

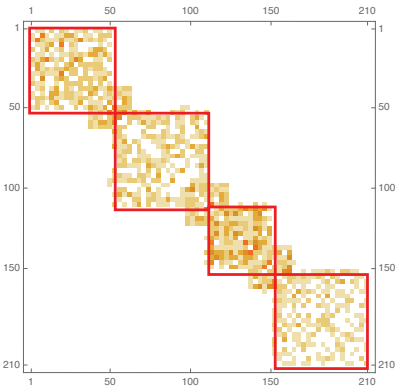
Imagine there is a network with a noisy sense of time. As depicted by the figure on the right, one can recognize different eras visually.
The idea is to penalize the weight of links based on their time distance. Therefore, the links between temporally neighbor nodes (nodes with close time stamp) will have higher weights than the others.



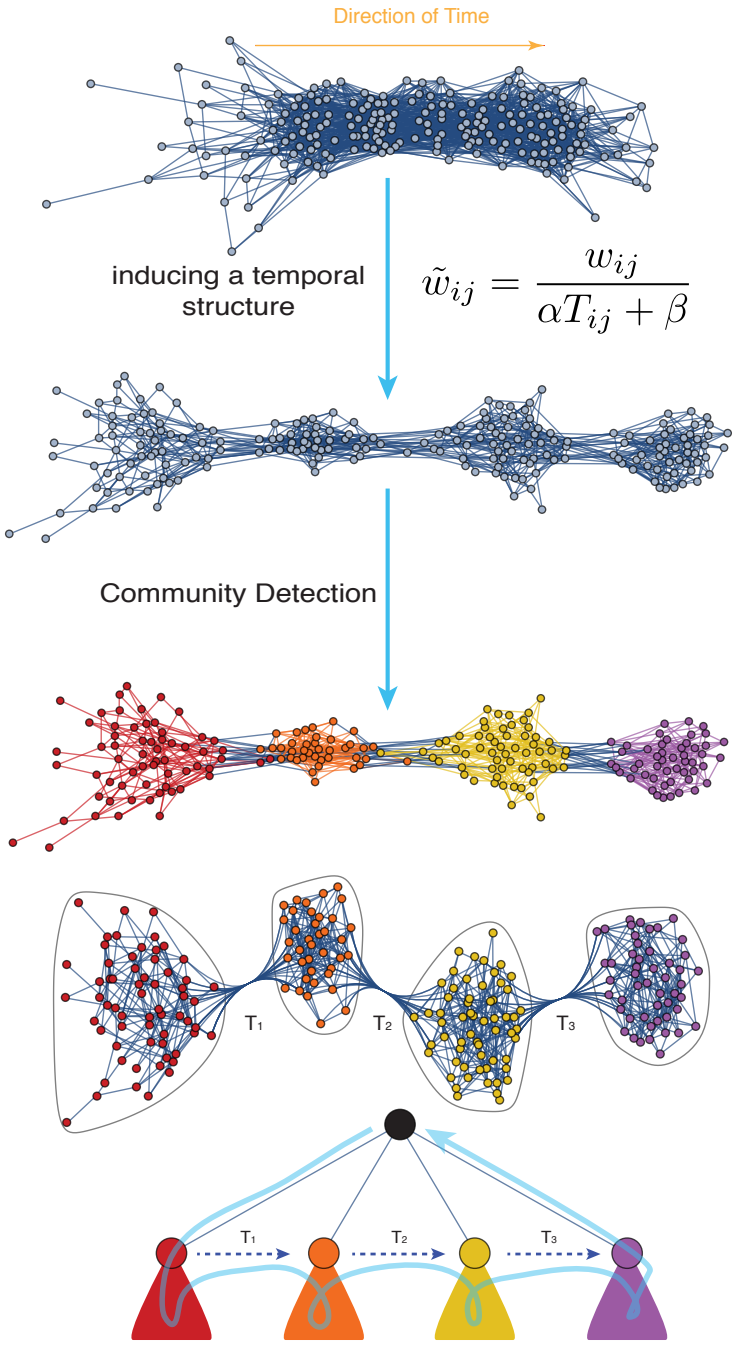
Even a simple linear penalization can improve the results dramatically. Here as we see the temporal structure of the network is accentuated after applying the penalization.



Now performing community detection gives a clean temporally ordered community structure.



Note that the first level jumps on the meta graph will correspond to paradigm shifts and can be interpreted specifically in the narrative.



Neo (Mostafa) Mohsenvand, MIT Media Lab
Facebook Fellowship Appendix E

References

1- Kurzweil, Ray. How to create a mind: The secret of human thought revealed. Penguin, 2012.

2- Steenkamp, Jan-Benedict EM, and Hans Baumgartner. "Development and cross-cultural validation of a short form of CSI as a measure of optimum stimulation level." International Journal of Research in Marketing 12.2 (1995): 97-104.

3- Hart, Sandra G. "NASA-task load index (NASA-TLX); 20 years later." Proceedings of the human factors and ergonomics society annual meeting. Vol. 50. No. 9. Sage Publications, 2006.