Using semantically-organized networks for the visualization of short texts

Jeongki Lim[1], Christina Boyce-Jacino[2], Dakota Murray[3], John F. Malloy[4], Ignacio Garnham[5], Pablo M. Flores[6], and Douglas Reckamp[7]

[1]Parsons School of Design, The New School

[2]Department of Social and Decision Sciences, Carnegie Mellon University

[3]School of Informatics, Computing, and Engineering, Indiana University Bloomington

[4]School of Earth and Space Exploration, Arizona State University

[5]School of Design Strategies, Parsons School of Design

[6]Department of Communication, University of California, Davis

[7]National War College, National Defense University

1 Abstract

Ethnographers, designers, and other qualitative researchers often analyze small-scale text content, such as interview transcripts. However, tools for text analysis are geared towards the study of large-scale textual corpora, and so have not been applied towards qualitative text analysis. Here, we propose a method using pre-trained *word2vec*—an algorithm for generating continuous vector representations of words from large corpora—for the visualization of small texts. Specifically, we use these embeddings to construct semantically-organized networks. These network visualizations are intuitive, tuneable, and offer a new means by which to gain insights about the semantic structure of a text. We demonstrate the utility of this approach by applying it to the text output of a design workshop hosted at the Santa Fe Institute in the Summer of 2019. The resulting semantically-organized network reveals the key structure of the topics and terms that resulted from the workshop which would not have been apparent from traditional means. We compare the semantically-organized network to spatial embeddings, and find advantage in the intuitive relational structure of the network. We discuss how this technique might be applied beyond design to the visualization of interviews, political speeches, online content analysis, and more.

2 Introduction

Text mining has proven essential to linguistic analysis across a variety of disciplines, powered by the increasing availability of large-scale linguistic corpora. However, even amid the present dominance data moment, many fields rely on analysis of smaller and more focused datasets [1]. Qualitative researchers often must make sense of interview and discourse transcriptions, short pieces of online content (blogs, social media posts, news stories), or text output from design methodology; standard approaches for text mining and visualizations do not allow for meaningful analysis of such small texts. Common approaches towards visualizing small text, such as word clouds, are difficult to interpret and fail to capture the semantic structure of text. Here, we propose a new approach towards visualizing small texts that uses word embeddings to construct a network representation of the semantic space of a text.

Since its publication in 2013, *word2vec* [12]—an algorithm for generating continuous vector representations of words (word embeddings) based on large-corpora—has proven useful for many applications. Specifically, they have been employed for information retrieval [3], content recommendation [13, 9, 14], cultural analytics [7, 4], and the study of general linguistic dynamics [5]. One benefit of *word2vec* is that word embeddings can be pre-trained on large corpora, and then easily applied towards other applications. Recognizing its interdisciplinary utility and ease of application, we examined the efficacy of *word2vec* for the visualization of the semantic structure of small-scale text. Specifically, we focus on its applicability to design research.

Often the design researchers make use of text data, but must rely on their intuition for anal-

ysis. Sometimes the text itself is not examined, yet is included a part of documentation of the research. Having approaches towards visualizing the semantic structure of interviews and output from design workshops allow for new ways of examining and presenting these texts, potentially multiplying their usefulness and the insights that can offer. Here, we examine the efficacy of a *word2vec*-based visualization method to design research by conducting an in-person design workshop and analyzing the text data with the word embedding technique.

3 Methods

The *semantic space* of a document is represented as a network created from word's pairwise similarities. These similarities are calculated using vectors resulting from *word2vec* [12], a method for learning continuous representations of words from a corpus of text—an *embedding*. In this embedding, a vector of numeric values is created where each value represents a feature of how that word is used in language. Cosine similarities between pairs of vectors roughly correspond to the degree of similarity between the contexts of each word. Here, the context of a word is defined as the set of words that appear before and after the word, up to a certain window size (typically between two and five). For example, consider the three sentences:

1. "We walked our **dog** at the park"
2. "I walked my **schnauzer** at the park"
3. "They rebooted the **computer** at the office"

Here, both ”dog” and ”schnauzer” share similar contexts, each including words like ”walked” and ”park”; these words would likely have high cosine similarity. However, words like ”rebooted” and ”office”, which are likely to appear in the context of ”computer” are less likely to appear in the context of ”dog”; these words would likely have low cosine similarity. By creating representations of words based on their contexts, *word2vec* makes it possible to measure the degree of *semantic similarity* between words. The *word2vec* algorithm learns these word representations from large corpora of text using a neural-network-based training method. Here, we use the pre-trained word vectors containing a vocabulary of nearly 3 million words trained on roughly 100 billion words (unigrams and phrases) from the Google News dataset, a large corpus of general English-language text.

Given a document, we remove numbers, punctuation, and convert all characters to lowercase. The text is then tokenized into unigrams and, optionally, phrases (tokens containing more than one word, i.e., ”New York City”). The vector representing each of these tokens are extracted from the pre-trained *word2vec* embeddings; tokens that do not appear in the embedding are excluded from further analysis. A matrix is created from the pairwise cosine similarities between each pair of vectors; For example, a document consisting of the words ”dog”, ”schnauzer” and ”computer” would result in the following matrix,

$$
\begin{array}{c} \\ dog \\ schnauzer \\ computer \end{array}
\begin{array}{c}
\begin{array}{ccc} dog & schnauzer & computer \end{array} \\
\begin{pmatrix} 1 & 0.71 & 0.12 \\ 0.71 & 1 & 0.06 \\ 0.12 & 0.06 & 1 \end{pmatrix}
\end{array}
$$

We consider the matrix of pairwise similarities as an undirected, weighted, and fully-connected network. We then use two approaches to extract the core structure of this network. The first is to find its maximum spanning tree. The spanning tree is defined as the sub-graph or tree that connects all the vertices (nodes) with the smallest number of edges possible. The maximum spanning tree is the spanning tree with the largest possible weights. The second approach is to use network backbone extraction [15]; this technique prunes edges that likely resulted from random connections between vertices. This approach depends on one parameter, α, which dictates the severity of the pruning. We visualize the resulting network using *webweb* [16], a tool built on d3.js that produces interactive network visualizations. A force-directed layout algorithm is used to assign positions to vertices.

4 Results and Discussion

We generated a semantically-organized network from the textual output of an informal design workshop held at the Santa Fe Institute during the Summer of 2019. The goal of this workshop was to investigate the cultural and social dimensions of three broad types of technologies: computers, lazers, and rocketry. Participants, consisting of about ten faculty and doctoral students affiliated

with the institute, were shown sets of images relating to each of these technologies. They were then asked to spend three minutes generating mental associations, writing down whatever terms or concepts they thought of when shown the images. These terms were collected and aggregated into a single document consisting of 383 unique terms.

Several versions of semantically-organized networks were created for these terms, with one shown in Fig 2. Many words appear in coherent clusters which can be manually grouped together (see Table 1. Each of these clusters highlights an aspect of the semantic space explored by the design workshop. Some clusters relate to the specifics of the technologies; for example, clusters D, G, and I contain words related to rocketry (e.g., ”capsule”, ”liftoff”, ”shuttle”), optics (e.g, ”lens”, ”prism”, ”wavelength”), and computing (e.g., ”processor”, ”linux”, ”tablet”), respectively. Other clusters instead speak to personal feelings of technology; for example, words in cluster A express a range of emotions surrounding these technologies (e.g., ”awe”, ”anticipation”, and ”tension”). Similarly, Cluster K illustrates the associations of these technologies with personal experience (e.g., ”travel”, ”shopping”, ”voyage”) and stories (e.g., ”writing”, ”fantasy”, ”ideas”). Some clusters instead speak to the cultural and social dimensions of these technologies; for example, cluster F contains words relating to the associates with these technologies to weaponry and warfare (e.g., ”blast”, ”fire”, ”gun”). Similarly, cluster H is small, but speaks to the relation of these technologies to empire and revolution. Cluster J illustrates terms that seem to relate to the practical consequences of computing, specifically in regards to teamwork, communication, connectivity, and control. The final cluster, C, contains words relating to areas of scientific research (e.g., ”physics”, ”scientist”, ”engineering), but also more abstract notions such as ”progress, ”advancement”, and

”innovation”.

The current standard for visualizing short text is the word cloud. However, standard word clouds plot frequency at the expense of semantic information and make it difficult to trace the structure of words in the text. For example, the word cloud for this design workshop (Fig 1), while demonstrating the prominence of words like ”space” and ”light”, fails to highlight how both these terms relate to broader sets of concepts that were prominent in the workshop. Semantically-organized word clouds have proven useful for visualization [17, 6], but they are less easily tuneable and obfuscate the inherent relational nature of language. The semantically-organized network makes clear how the terms and concepts explored during the design workshop relate to one another, and by grouping terms together, distil the semantic space to a smaller set of coherent concepts (table 1).

Parameters Semantically-organized networks can be tuned to best fit the particular visualization context. Here, we outline two approaches to extract the semantically-organized network, one using a maximum-spanning tree approach, and another using network backbone extraction. The maximum-spanning tree (Fig 3) plots words as leaves on a branching tree structure. Under this approach, word similarity can be assessed based on their relative positions on each branch, and hierarchical clusters can be easily identified. More specific words will tend to appear on the outer leaves, wheres more general words will tend to appear near the center. This approach also ensures that the main component of the network includes all words. However, this visualization makes it difficult to understand the complex relationships between sets of words and and makes

Figure 1: **Word Cloud** showing the frequency of words generated in the SFI workshop. The more often a word was generated by a participant, the larger it appears in the cloud.

semantically-distant relationships indistinguishable form more similar ones. Network-backbone extraction provides a tunable alternative which can be used to understand the core structures of the network. By adjusting the parameter, α, the user can adjust the severity of the edge pruning procedure, making a more disconnected graph (as for $\alpha = 0.01$, Fig 3) or a more fully-connected and dense graph (as for $alpha > 0.05$). We recommend using both maximum-spanning tree and network backbone extraction when exploring new data. When using backbone extraction, a range of parameter values can also be used in order to assess the stability of word clusters.

Limitations This technique is subject to several limitations. First, *word2vec* embeds words based on their contexts across a large corpus of English language and will not represent specific contexts. Words with multiple meanings may be visualized in non-intuitive ways. This is most common

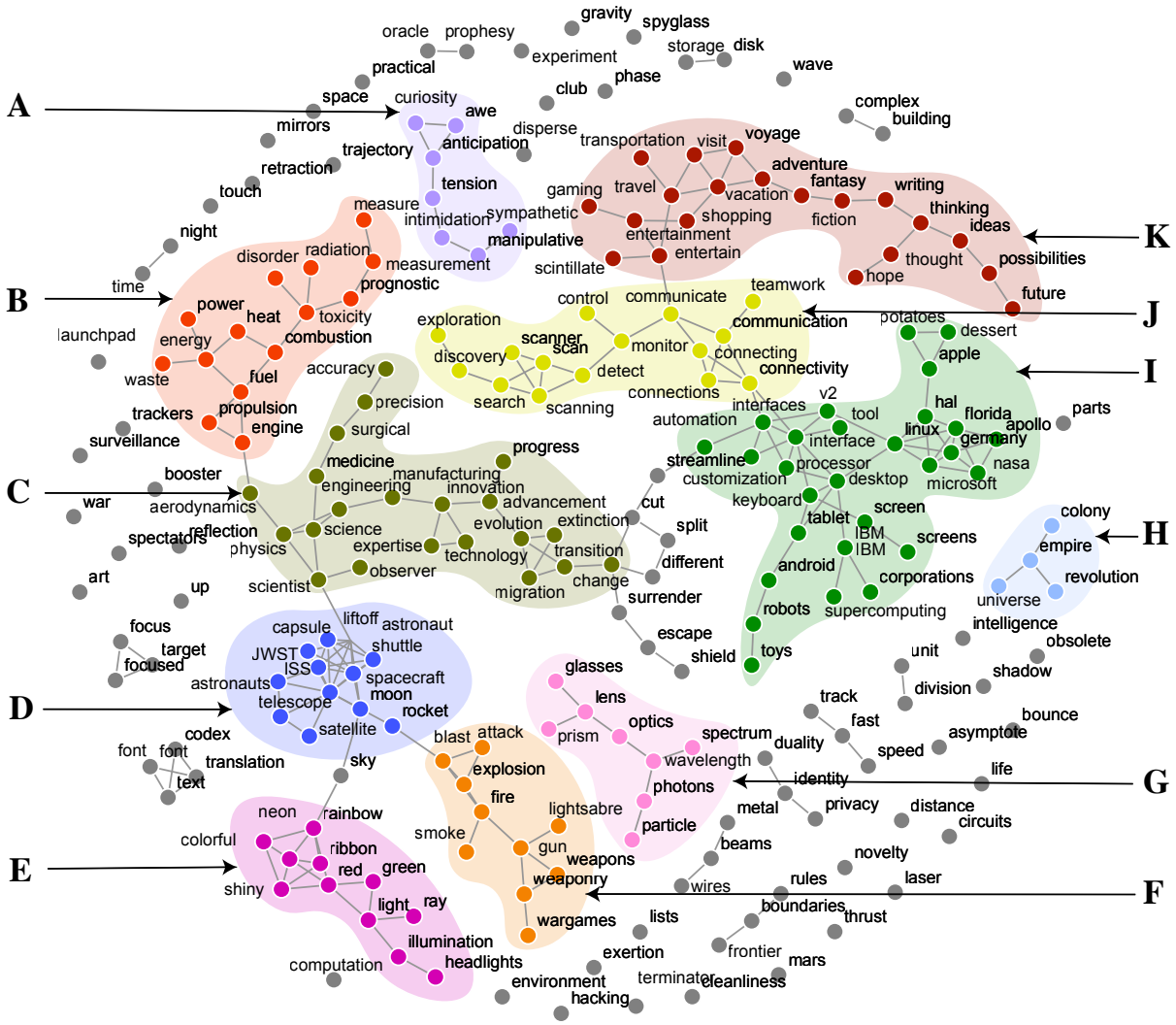

Figure 2: **Semantic-organized network of design workshop.**. Network backbone extracted from the matrix of pairwise cosine similarities between terms embeddings of terms from the workshop hosted at the Santa Fe Institute. A parameter of $\alpha = 0.025$ was used to for backbone extraction. Every term from the workshop that appeared in the *word2vec* pre-trained embeddings is mapped to a node. Edges indicate a high level of semantic similarity between words. Meaningful groupings of terms were manually identified, colored, and labeled from **A** to **K**.

for proper nouns that share the same spelling as other common words. For example, the word "curiosity" in cluster A of Fig 2 likely refers to Curiosity the Mars Rover, rather than the emotion. Similarity, the word "Apple" in cluster I could refer to company, rather then the fruit. Fortunately,

word vectors can be trivially replaced with alternatives, and so context-specific word vectors can be used when appropriate. Secondly, the original word vectors were trained on Google's News corpus which is not a representative sample of the English language, and so may not contain certain words, or may produce non-representative word vectors. These vectors were also trained on past text, and so they cannot accommodate new words or changing definitions. Finally, we note that the visualizations we create and clusters we identify are subjective and prone to misinterpretation; therefore we the primary utility of this approach as exploratory.

Applications Here we have shown how this method might be applied towards the visualization and analysis of text data resulting from a design workshop or a similar communal word-generating scenario. However, this technique is trivially generalizable and has diverse applications. For example, design researchers and ethnographers often make use of interviews to gain understandings of people and their experiences— building a semantic-organized network from these interview offers a novel means of generating insights. For example, Fig 4 illustrates how, in an example interview with a mechanical engineering, that the interviewee explored topics such as education, geography, and the specifics of the work. Political scientists and historians may employ this technique to visualize and understand variations in political speeches. For example, Fig 5 illustrates how Barack Obama and Donald Trump each spoke of the political and social landscape of the U.S. in 2016 and 2018, respectively. Whereas Obama spoke of individual topics relating to the opioid crisis, immigration, and politics, Trump instead spoke broadly on recent natural disasters and terrorism, and used far more emotive terms such as "hardship", "daring", and "terrible". Beyond the examples presented here, this technique can be applied to any problem domain involving small texts,

including the visualization of blogs, online user reviews, sections of books or novels, and more.

Other embedding methodology One particular feature of the embeddings generated by word2vec is that they are high dimensional; each word has a 300-dimensional vector of features. One advantage of the network based visualization and pruning described above, is that it effectively reduces the dimensionality of the word vectors down to just the degree of similarity between two words. However, this representation need not necessarily be a network. There are a variety of dimensionality-reduction techniques that can be applied to generate spatial or clustered, rather than relational, representations of words.

One potential method is to simply work with the word vectors themselves, extracting structure from a set of words based on emergent clusters and semantic relationships. What we want to learn is what words sit close together in semantic space, that is, are there areas of high density in our text. The class of algorithms to use here are density based clustering algorithms. In the first example presented here (see Fig 6) we use k-means clustering [8] to probabilistically assign words to one of eight clusters. We then use a dimensionality-reduction technique called t-SNE [10] to project our word vectors down to two dimensions, and color the word points according to their cluster membership. Because t-SNE does not faithfully preserve local density or topology, we see that words which have, by our k-means clustering algorithm, been assigned to the same cluster, appear in the 2-dimensional rendering, far apart. The preservation of local structure is a key advantage of semantically-organized networks.

In a similar vein, we can apply different clustering and dimensionality reduction techniques

to try to better preserve the natural semantic clustering of our words. One method we apply is a combination of a uniform manifold approximation and projection (UMAP, [11]) and hierarchical clustering algorithm, HDBSCAN [2]. The first step of this method is to use UMAP to reduce the dimensions in our word vectors to a small enough number that density-based clustering methods are maximally effective. As in t-SNE, UMAP seeks to project high dimensional vectors down to fewer dimensions, however unlike in t-SNE, we can now define how many dimensions we wish to project on to. For ease of comparison, we project down to two dimensions. Using these lower-dimensional values, we then apply our clustering algorithm. Unlike k-means clustering, HDBSCAN can refuse to classify, or assign to a cluster, data points which it instead marks as noise. It also does not require *a priori* specification of a number of clusters to discover; the method used here discovers four clusters, and opts to not classify a significant part of the data (See Fig 7. While this approach is more informative than t-SNE, it tends to create large centralized and noisy clusters and obfuscates to what extent clusters are related.

5 Conclusion

Here we have introduced semantically-organized networks, a new approach to visualizing small documents using word embeddings to organize words in an intuitive graph structure. By leveraging relational structure, this approach can be more informative than simple projects of word embeddings and semantically-organized word clouds. Moreover, this approach is tunable, easily generalizable, and can easily be modified with language- or context-specific word embeddings. In this study we have demonstrated how this approach might be used by designers to gain insights

from a design workshop. However, we have also outlined other potential use cases, such as for the analysis of interviews, speech, and more. Small documents present unique challenges for analysis. With semantically-organized networks, we offer a new and valuable means of approaching and gaining insight from these documents.

References

[1] Christine L. Borgman. *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge, Massachusetts: The MIT Press, Jan. 2015. ISBN: 978-0-262-02856-1.

[2] Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. "Density-Based Clustering Based on Hierarchical Density Estimates". In: *Advances in Knowledge Discovery and Data Mining*. Springer Berlin Heidelberg, 2013, pp. 160–172. DOI: 10.1007/978-3-642-37456-2_14. URL: https://doi.org/10.1007/978-3-642-37456-2_14.

[3] Debasis Ganguly et al. "Word Embedding Based Generalized Language Model for Information Retrieval". In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '15. event-place: Santiago, Chile. New York, NY, USA: ACM, 2015, pp. 795–798. ISBN: 978-1-4503-3621-5. DOI: 10.1145/2766462.2767780.

[4] Nikhil Garg et al. "Word embeddings quantify 100 years of gender and ethnic stereotypes". In: *Proceedings of the National Academy of Sciences* 115.16 (Apr. 2018), E3635–E3644. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1720347115. URL: https://www.pnas.org/content/115/16/E3635.

[5] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. “Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change”. In: *arXiv:1605.09096 [cs]* (Oct. 2018). arXiv: 1605.09096.

[6] Marti Hearst et al. “An Evaluation of Semantically Grouped Word Cloud Designs”. In: *IEEE transactions on visualization and computer graphics* (Mar. 2019). ISSN: 1941-0506. DOI: 10.1109/TVCG.2019.2904683.

[7] Austin C. Kozlowski, Matt Taddy, and James A. Evans. “The Geometry of Culture: Analyzing Meaning through Word Embeddings”. In: *arXiv:1803.09288 [cs]* (Mar. 2018). arXiv: 1803.09288.

[8] Aristidis Likas, Nikos Vlassis, and Jakob J. Verbeek. “The global k-means clustering algorithm”. In: *Pattern Recognition* 36.2 (Feb. 2003), pp. 451–461. DOI: 10.1016/s0031-3203(02)00060-2.

[9] Hualong Ma et al. “Course recommendation based on semantic similarity analysis”. In: *2017 3rd IEEE International Conference on Control Science and Systems Engineering (ICCSSE)*. IEEE, Aug. 2017. DOI: 10.1109/ccsse.2017.8088011.

[10] Laurens van der Maaten and Geoffrey Hinton. “Visualizing Data using t-SNE”. In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605.

[11] Leland McInnes, John Healy, and James Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. 2018. arXiv: 1802.03426 [stat.ML].

[12] Tomas Mikolov et al. “Distributed Representations of Words and Phrases and Their Compositionality”. In: *Proceedings of the 26th International Conference on Neural Information*

Processing Systems - Volume 2. NIPS'13. USA: Curran Associates Inc., 2013, pp. 3111–3119.

[13] Cataldo Musto et al. "Word Embedding Techniques for Content-based Recommender Systems: An Empirical Evaluation." In: *RecSys Posters*. Ed. by Pablo Castells. Vol. 1441. CEUR Workshop Proceedings. CEUR-WS.org, 2015.

[14] Makbule Gulcin Ozsoy. *From Word Embeddings to Item Recommendation*. 2016. arXiv: 1601.01356 [cs.LG].

[15] M. Ángeles Serrano, Marián Boguñá, and Alessandro Vespignani. "Extracting the multiscale backbone of complex weighted networks". In: *Proceedings of the National Academy of Sciences* 106.16 (Apr. 2009), pp. 6483–6488. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0808904106. (Visited on 05/01/2018).

[16] K. Wapman and Daniel Larremore. "webweb: a tool for creating, displaying, and sharing interactive network visualizations on the web". In: *Journal of Open Source Software* 4.40 (Aug. 12, 2019), p. 1458. ISSN: 2475-9066. DOI: 10.21105/joss.01458.

[17] Jin Xu, Yubo Tao, and Hai Lin. "Semantic word cloud generation based on word embeddings". In: *2016 IEEE Pacific Visualization Symposium (PacificVis)*. ISSN: 2165-8773. Apr. 2016, pp. 239–243. DOI: 10.1109/PACIFICVIS.2016.7465278.

Table 1: **Cluster labels and names.** Labels correspond to groups of terms marked in Fig 2. Descriptive names were manually assigned based on interpretation of terms within each group.

Label	Descriptive Name	Example words
A	Emotive	awe, anticipation, tension
B	Engineering	engine, radiation, measure
C	Knowledge development	science, innovation, progress
D	Space science	JWST, rocket, moon
E	Spectrum	rainbow, neon, light
F	Weaponry	gun, blast, lightsabre
G	Optics	optics, lens, photons
H	Expansion	universe, empire, revolution
I	Computing	processor, tablet, linux
J	Experience and stories	travel, shopping, writing

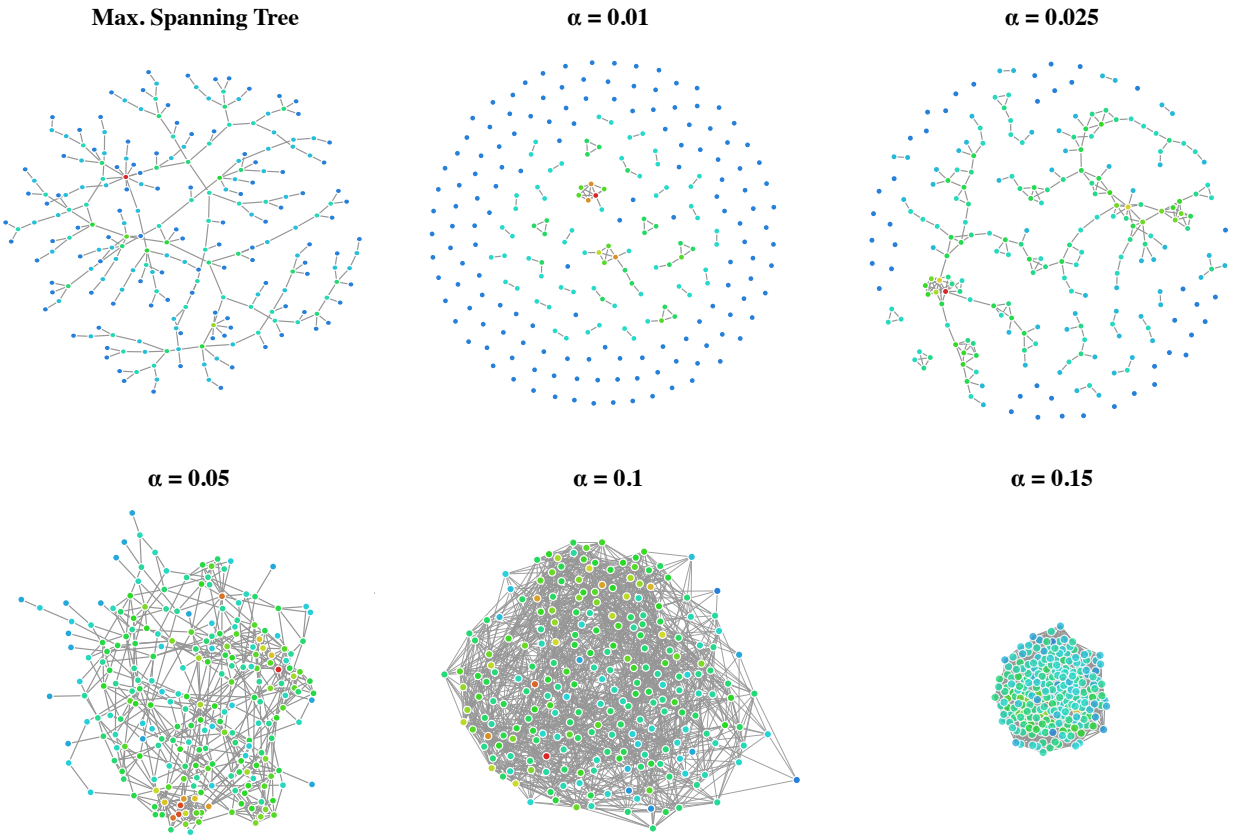

Figure 3: **The effect of network extraction technique and parameter.** Nodes are colored by their degree relative to each network; red indicates a high degree whereas blue a low degree. Edges shown are those that remain after extracting the network from the matrix of pairwise similarities. Pictured top left is the network extracted using the maximum-spanning tree technique. All others resulted from network backbone extraction with a varying α, which dictates the severity of the pruning.

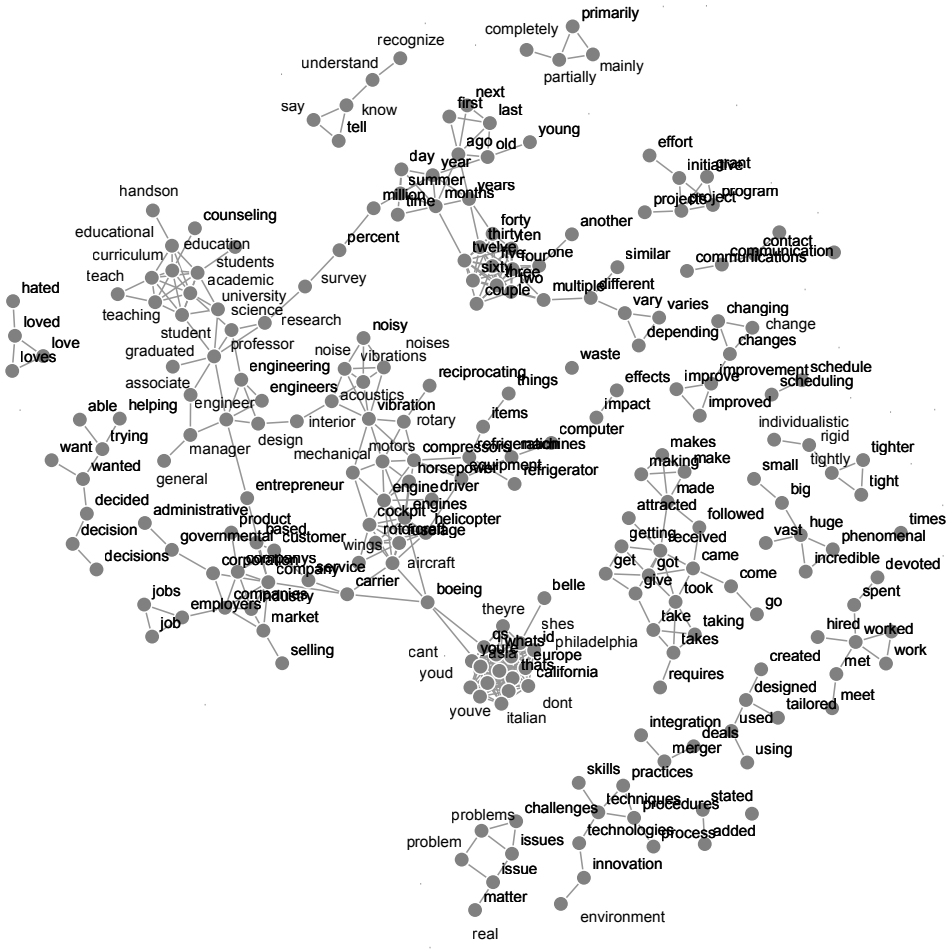

Figure 4: Visualization of interview transcript given by a mechanical engineer regarding their work and experiences. This transcript can be found at *http://www.u.arizona.edu/~kimmehea/purdue/421/exampleinterview.html*. Only the main connected components are shown. Network backbone extraction was used with $\alpha = 0.01$.

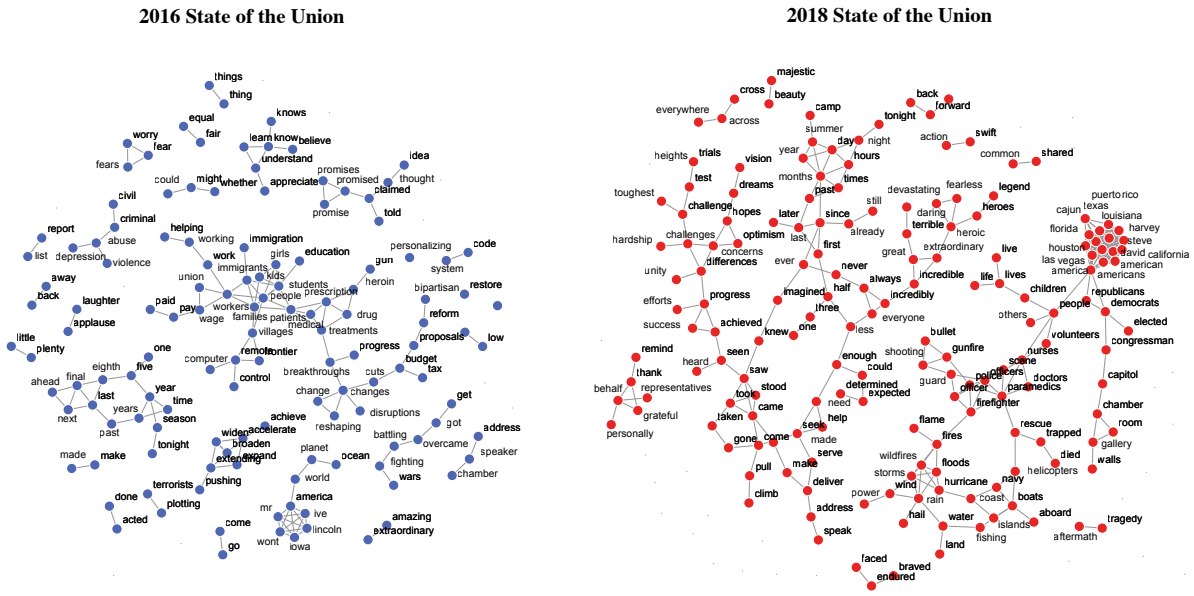

Figure 5: Visualization first 500 words of the state of the 2016 State of the Union Address given by Barack Obama (left), and the 2018 State of the Union Address given by Donald Trump (right). Only the main connected components are shown. Both were created using network backbone with $\alpha - 0.025$.

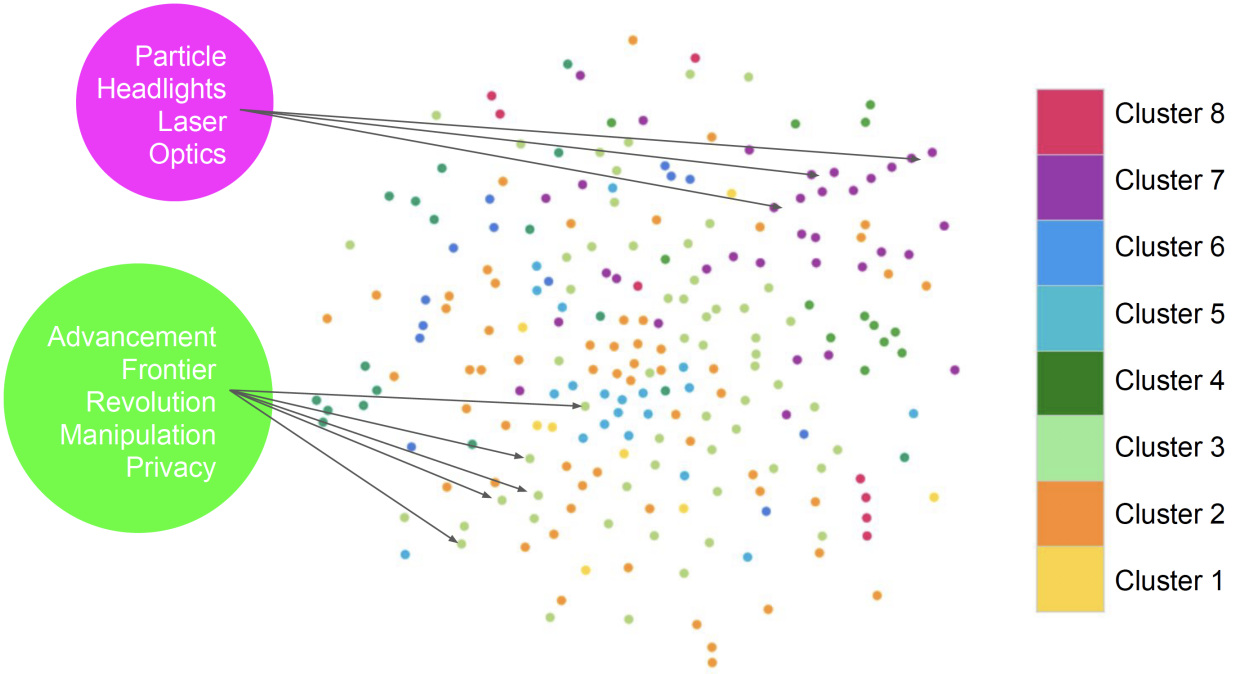

Figure 6: Example visualization and clustering of word vectors of terms generated by the design workshop. Labels report words from two selected clusters, one highlighting a cluster of more ”concrete” terms, and another highlighting different, more ”abstract” terminology used to describe the technological images shown.

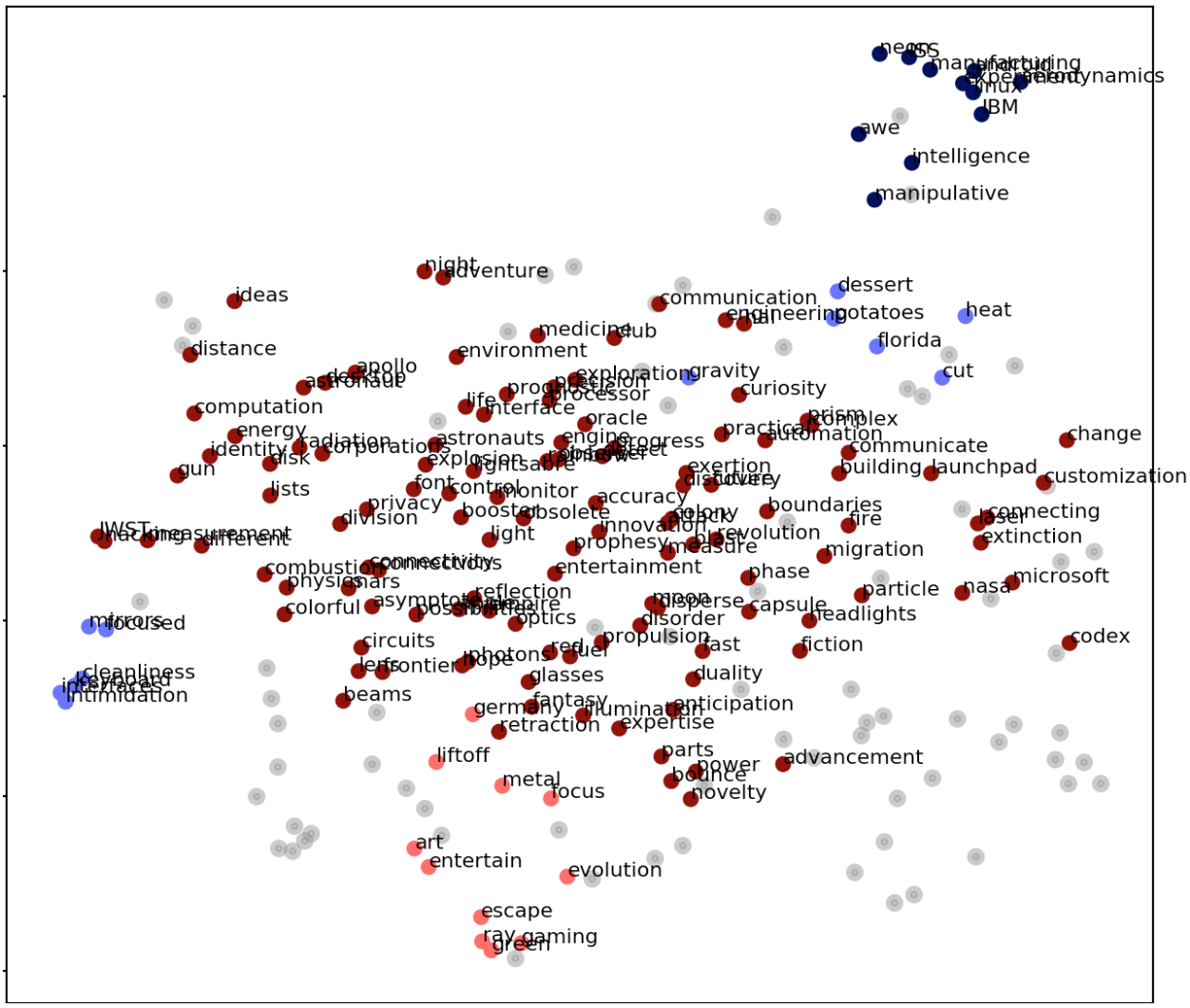

Figure 7: Visualization of workshop generated terms using UMAP embeddings.