Big Signal Processing for Multi-Aspect Data Mining

Evangelos E. Papalexakis, Carnegie Mellon University
http://www.cs.cmu.edu/~epapalex

What does a social graph between people who *call* each other look like? How does it differ from one where people *instant-message* or *e-mail* each other? Social interactions, along with many other real-word processes and phenomena, have different *aspects*, such as the means of communication. In the above example, the activity of people calling each other will likely differ from the activity of people instant-messaging each other. Nevertheless, each aspect of the interaction is a signature of the same underlying social phenomenon: formation of social ties and communities. Taking into account all aspects of social interaction results in more accurate social models (e.g, communities). The main thesis of my work is that many real-world problems, such as the aforementioned, benefit from jointly modeling and analyzing the multi-aspect data associated with the underlying phenomenon we seek to uncover. In conclusion, I focus on scalable and interpretable algorithms for mining big multi-aspect data by **bridging Signal Processing and Data Science for real-world applications**.

THESIS WORK

My thesis work is broken down to *algorithms* with contributions mostly in tensor analysis as well as other fields such as control system identification [10], and multi-aspect data mining *applications*.

A) Algorithms

The primary computational tool in my work is *tensor decomposition*. Tensors are multi-dimensional matrices, where each *aspect* of the data is mapped to one of the dimensions or *modes*. In order to analyze a tensor, we compute a decomposition or factorization (henceforth we use the terms interchangeably), which gives a low-dimensional *embedding* of all the aspects. I focused on the Canonical or PARAFAC decomposition, which decomposes the tensor into a sum of outer products of *latent factors* (see also Figure 1):

$$\underline{\mathbf{X}} \approx \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r$$

where $\circ$ denotes outer product, i.e., $[\mathbf{a} \circ \mathbf{b} \circ \mathbf{c}](i,j,k) = \mathbf{a}(i)\mathbf{b}(j)\mathbf{c}(k)$. Informally, each latent factor is a co-cluster of the aspects of the tensor. The advantages of this decomposition over other existing ones are interpretability (each factor corresponds to a co-cluster), and strong uniqueness guarantees for the latent factors.Tensor decompositions are very powerful tools, with numerous applications (see details below). There is increasing interest in their application to Big Data problems, both from academia and industry. However, algorithmically, there exist challenges which limit their applicability to real-world, big data problems, pertaining to scalability and quality assessment of the results. *Below I outline how my work addresses those challenges, towards a broad adoption of tensor decompositions in big data science.*

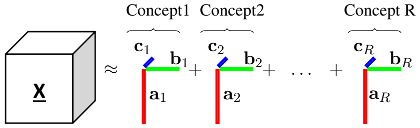
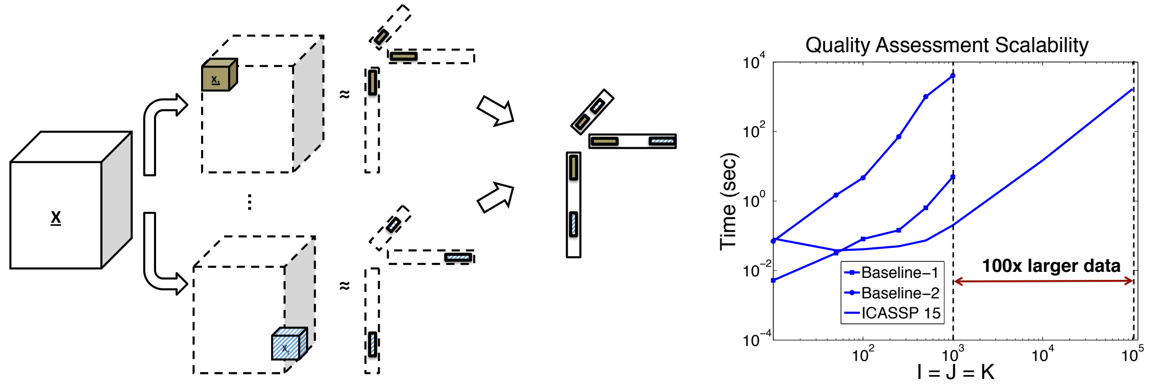


Figure 1: Canonical or PARAFAC decomposition into sum of R rank-one components. Each component is a *latent concept* or a *co-cluster*.

A1) Parallel, and Scalable Tensor Decompositions

Consider a multi-aspect tensor dataset that is too big to fit in the main memory of a single machine. The data may have large "ambient" dimension (e.g., a social network can have billions of users), however the observed interactions are very sparse, resulting in extremely sparse data. This data sparsity can be exploited for efficiency. In [9] we formalize the above statement by proposing the concept of a **triple-sparse algorithm** where 1) the input data are sparse, 2) the intermediate data that the algorithm is manipulating or creating are sparse, and 3) the output is sparse. Sparsity in the intermediate data is crucial for scalability. In [15] we show that the intermediate data created by a tensor decomposition algorithm designed for dense data can be many orders of magnitude larger than the original data, rendering the analysis prohibitive. Sparsity in the results is a great advantage, both in terms of storage but most importantly in terms of interpretability, since sparse models are easier for humans to inspect. *None of the existing state of the art algorithms, before[9], fulfilled all three requirements for sparsity.*

In [9] we propose ParCube, the first triple-sparse, parallel algorithm for tensor decomposition. Figure 2(a) depicts a high-level overview of ParCube. Suppose that we computed a weight of how important every row, column, and "fiber" (the third mode index) of the tensor is. Given that weight, ParCube takes biased samples of rows, columns, and fibers, extracting a small tensor from the full data. This is done repeatedly with each sub-tensor effectively explores different parts of the data. Subsequently, ParCube decomposes all those sub-tensors *in parallel* generating partial results. Finally, ParCube merges the partial results ensuring that partial results corresponding to the same latent component are merged together. The power behind ParCube is that, even though the tensor itself might not fit in memory, we can choose the sub-tensors appropriately so that they fit in memory, and we can compensate by extracting many independent sub-tensors. ParCube converges to the same level of sparsity as [13] (the first tensor decomposition with latent sparsity) and furthermore ParCube's approximation error converges to the one of the full decomposition (in cases where we are able to run the full decomposition). This demonstrates that ParCube's sparsity maintains the useful information in the data. In [11], we extend the idea of [9], introducing Turbo-SMT, for the case of Coupled Matrix-Tensor Factorization (CMTF), where a tensor and a matrix share one of the aspects of the data, achieving up to **200 times faster** execution with comparable accuracy to the baseline, on a single machine. Subsequently, in [17] we propose ParaComp, a novel parallel architecture for tensor decomposition in similar spirit as ParCube. Instead of sampling, ParaComp uses random



(a) The main idea behind PARCUBE: Using biased sampling, extract small representative sub-sampled tensors, decompose them in parallel, and carefully merge the final results into a set of sparse latent factors.

(b) Computing the decomposition quality for tensors for two orders of magnitude larger tensor than the state of the art (I, J, K are the tensor dimensions).

Figure 2

projections to *compress* the original tensor to multiple smaller tensors. Thanks to compression, in [17] we prove that PARACOMP can guarantee the *uniqueness* of the results (cf. [17] for exact bounds and conditions). This is a very strong guarantee on the correctness and quality of the result.

In addition to [9, 11, 17], which introduce a novel paradigm for parallelizing and scaling up tensor decomposition, in [15] we developed the first scalable algorithm for tensor decompositions on Hadoop which was able to decompose problems larger by at least **two orders of magnitude** than the state of the art. Subsequently [4] we developed a Distributed Stochastic Gradient Descent method for Hadoop that scales to billions of parameters.

A2) Unsupervised Quality Assessment of Tensor Decompositions

Real-world exploratory analysis of multi-aspect data is, to a great extent, *unsupervised*. Obtaining ground truth is a very expensive and slow process, or in the worst case impossible; for instance, in *Neurosemantics* where we research how language is represented in the brain, most of the subject matter is uncharted territory and our analysis drives the exploration. Nevertheless, we would like to assess the quality of our results in absence of ground truth. There is a very effective heuristic in Signal Processing and Chemometrics literature by the name of "Core Consistency Diagnostic" (cf. Bro and Kiers, Journal of Chemometrics, 2003) which assigns a "quality" number to a given tensor decomposition and gives information about the data being inappropriate for such analysis, or the number of latent factors being incorrect. However, this diagnostic has been specifically designed for fully dense and small datasets, and is not able to scale to large and sparse data. In [8], exploiting sparsity, we introduce a *provably exact algorithm* that operates on at least **two orders of magnitude larger data** than the state of the art (as shown in Figure 2(b)), which enables quality assessment on large real datasets for the first time.

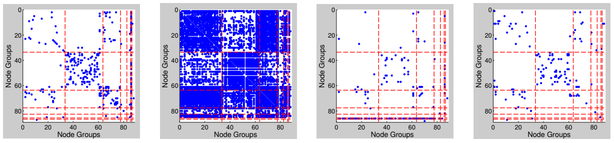
Impact - Algorithms

- PARCUBE [9] is the most cited paper of ECML-PKDD 2012 with 46 citations at the time of writing, whereas the median number of citations for ECML-PKDD 2012 is 5. Additionally, PARCUBE has already been downloaded more than 80 times by universities and organizations from 23 countries.
- TURBO-SMT [11] was selected as one of the best papers of SDM 2014, and will appear in a special issue of the Statistical Analysis and Data Mining journal.
- PARACOMP [17] has appeared in the prestigious IEEE Signal Processing Magazine.

B) Applications

My work has focused on: 1) *Multi-Aspect Graph Mining*, 2) *Neurosemantics*, and 3) *Knowledge on the Web*.

B1) Multi-Aspect Graph Mining



(a) `calls` (b) `proximity` (c) `sms` (d) `friends`

Figure 3: Results on the four views of the REALITYMINING multi-graph. Red dashed lines outline the clustering found by GRAPHFUSE.

In [5] we introduce GRAPHFUSE, a tensor based community detection algorithm which uses different aspects of social interaction and outperforms state of the art in community extraction accuracy. Figure 3 shows GRAPHFUSE at work, identifying communities in REALITYMINING, a real dataset of multi-aspect interactions between students and faculty at MIT. The communities are consistent across aspects and agree with ground truth. Another aspect of a social network is *time*. In [3] we introduce COM2, a tensor based *temporal community* detection algorithm, which identifies social communities and their behavior over time. In [7] we consider *language* as another aspect, where we identify topical and temporal communities in a discussion forum of Turkish immigrants in the Netherlands, and in [12] we consider *location* (which has become extremely pervasive recently) analyzing data from Foursquare, identifying spatial and temporal patterns of users' activity in Foursquare. My work is not necessarily restricted to social networks, e.g., I have also analyzed multi-aspect computer network graphs, detecting *anomalies* and *network attacks*[6, 16].

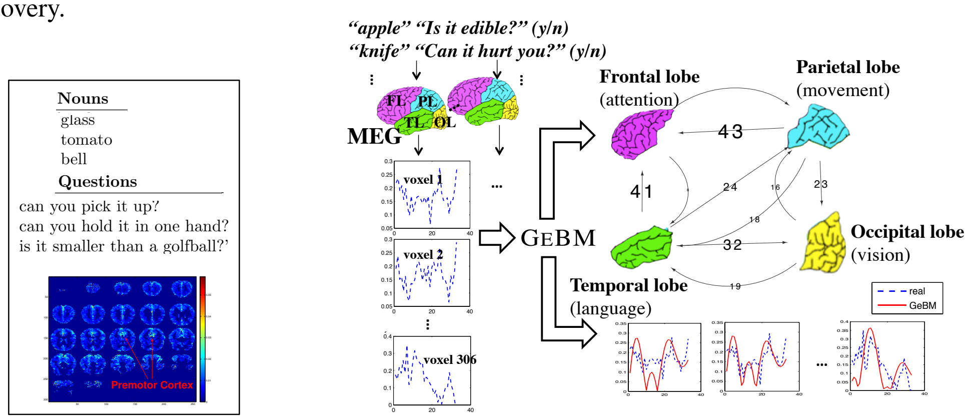
Impact - Multi-Aspect Graph Mining

- COM2 [3] won the best student paper award at PAKDD 2014.
- GRAPHFUSE [5] has been downloaded more than 80 times from 21 countries.
- Our work in [16] is deployed by the Institute for Information Industry in Taiwan, detecting real network intrusion attempts.
- Our work in [6] was selected, as one of the best papers of ASONAM 2012, to appear in Springer's Encyclopedia for Social Network Analysis and Mining.

B2) Neurosemantics

How is knowledge represented in the human brain? Which regions have high activity and information flow, when a concept such as "food" is shown to a human subject? Do all human subjects' brains behave similarly in this context? Consider the following experimental setting, where multiple human subjects are shown a set of concrete English nouns (e.g. "dog", "tomato" etc), and we measure each person's brain activity using various techniques (e.g, fMRI or MEG). In this experiments, human subjects, semantic stimuli (i.e., the nouns), and measurement methods are all different aspects of the same underlying phenomenon: the mechanisms that the brain uses to process language.

In [11], we seek to identify coherent regions of the brain that are activated for a semantically coherent set of stimuli. To that end we combine fMRI measurements with semantic features (in the form of simple questions, such as *Can you pick it up?*) for the same set of nouns, which provide useful information to the decomposition which might be missing from the fMRI data, as well as constitute a human readable description of the semantic context of each latent group. A very exciting example of our results can be seen in Figure 4(a), where all the nouns in the "cluster" are small objects, the corresponding questions reflect holding or picking such objects up, and most importantly, the brain region that was highly active for this set of nouns and questions was the *premotor cortex*, which is associated with holding or picking small items up. This result is **entirely unsupervised** and agrees with Neuroscience. This gives us confidence that the same technique can be used in more complex tasks (cf. future research) and drive neuroscientific discovery.



(a) The premotor cortex, having high activation here, is associated with motions such as holding and picking small items up.

(b) Given MEG measurements (time series of the magnetic activity of a particular region of the brain), GEBM learns a graph between physical brain regions. Furthermore, GEBM simulates the real MEG activity very accurately.

Figure 4: Overview of results of the Neurosemantics application.

In a similar experimental setting, where the human subjects are also asked to answer a simple yes/no question about the noun they are reading, in [10] we seek to discover the *functional connectivity* of the brain for the particular task. Functional connectivity is an information flow graph between different regions of the brain, indicating high degree of interaction between (not necessarily physically connected) regions while the person is reading the noun and answering the question. [10] we propose GEBM, a novel model for the functional connectivity which views the brain as a *control system* and we propose a sparse system identification algorithm which solves the model. Figure 4(b) shows an overview of GEBM: given MEG measurements (time series of the magnetic activity of a particular region of the brain), we learn a model that describes a graph between physical brain regions, and simulates real MEG activity very accurately.

Impact - Neurosemantics

- GEBM [10] is taught in class `CptS 595` at Washington State University.
- Our work in [11] was selected as one of the best papers of SDM 2014

B3) Knowledge on the Web

Knowledge on the Web has multiple aspects: real-world entities such as *Barack Obama* and *USA* are usually linked in multiple ways, e.g., *is president of*, *was born in*, and *lives in*. Modeling those multi-aspect relations a tensor, and computing a low rank decomposition of the data, results in *embeddings* of those entities in a lower dimension, which can help discover semantically and contextually similar entities, as well as discover missing links. In [9] we discover semantically similar noun-phrases in a Knowledge Base coming from the *Read the Web* project at CMU: `http://rtw.ml.cmu.edu/rtw/`. *Language* is another aspect: many web-pages have parallel content in different languages, however, some languages have higher representation than others. How can we learn a high quality joint latent representation of entities and words, where we combine information from all languages? In [14] we introduce the notion of *translation-invariant word embeddings* where we compute multi-lingual embeddings, forcing translations to be "close" in the embedding space. Our approach outperforms the state of the art.

Yet another aspect of an real-world entity on the web is the set of *search engine results* for that entity, which is the biased view of each search engine, as a result of their crawling, indexing, ranking, and potential personalization algorithms, for that query. In [1] we introduce TENSORCOMPARE, a tool which measures the overlap in the results of different search engines. We conduct a case study on Google and Bing, finding high overlap. Given this high overlap, how can we use different signals, potentially coming from social media, in order to provide diverse and useful results? In [2] we follow-up designing a Twitter based web search engine, using tweet popularity as a ranking function, which does exactly that. This result has huge potential for the future of web search, paving the way for the use of social signals in the determination and ranking of results.

Impact - Knowledge on the Web

- In [14] we are the first to propose the concept of translation-invariant word embeddings.
- Our work in [1] was selected to appear in a special issue of the Journal of Web Science, as one of the best papers in the Web Science Track of WWW 2015.

FUTURE RESEARCH DIRECTIONS

As more field sciences are incorporating information technologies (with Computational Neuroscience being a prime example from a long list of disciplines), the need for scalable, efficient, and interpretable multi-aspect data mining algorithms will only increase.

1) Long-Term Vision: Big Signal Processing for Data Science

The process of extracting useful and novel knowledge from big data in order to drive scientific discovery is the holy grail of data science. Consider the case where we view the knowledge extraction process through a signal processing lens: suppose that a transmitter (a physical or social phenomenon) is generating signals which describe aspects of the underlying phenomenon. An example signal is a time-evolving graph (a time-series of graphs) which can be represented as a tensor. The receiver (the data scientist in our case), combines all those signals with ultimate goal the reconstruction (and general understanding) of their generative process. The "communication channel" wherein the data are transmitted can play the role of the measurement process where loss of data occurs, and thus we have to account the channel estimation into our analysis, in order to reverse its effect. We may consider that the best way to transmit all the data to the receiver is to find the best compression or dimensionality reduction of those signals (e.g., *Compressed Sensing*). There may be more than one transmitters (if we consider a setting where the data are distributed across data centers) and multiple receivers, in which case privacy considerations come into play. In my work so far I have established two connections between Signal Processing and Data Science (Tensor Decompositions & System Identification), contributing new results in both communities and have already had significant impact, demonstrated, for instance, by the amount of researchers using and extending my work. These two aforementioned connections are but instances of a vast landscape of opportunities for cross-pollination which will advance the state of the art of both fields and drive scientific discovery. I envision my future work as a three-way bridge between Signal Processing, Data Science, and high impact real-world applications.

2) Mid-Term Research Plan

With respect to my shorter term plan (first 3-5 years), I provide a more detailed outline of the thrusts and challenges that I am planning to tackle, both regarding algorithms and multi-aspect data applications.

Algorithms: Triple-Sparse Multi-Aspect Data Exploration & Analysis with Quality Assessment

In addition to improving the state of the art for tensor decompositions, I plan to explore alternatives for multi-aspect data modeling and representation learning, which can be used with tensor decompositions. Some of the challenges are:
Algorithms, models, and problem formulation: What is the appropriate (factorization) model? Are there any noisy aspects that may hurt performance? As the ambient dimensions of the data grow, and the number of aspects increases, data become much sparser. Sparsity, as we saw is a blessing when it comes to scalability, however it can also be a curse when it comes to detecting meaningful relatively dense patterns in subspaces within the data.
Scalability & Efficiency: Data size will continue to grow and we need faster and more scalable algorithms to handle the size and complexity of the data. This will involve adjusting existing algorithms or proposing new algorithms that adhere to the *triple-sparse* paradigm. Furthermore, the particular choice of a distributed system can make a huge difference depending on the application. For instance, Map/Reduce works well in batch tasks, whereas it has well known weaknesses in iterative algorithms. Other systems, e.g., Spark could be more appropriate for such tasks, and in the future I plan to research the capabilities of current high-end distributed systems, in relation to the algorithms I develop.
(Unsupervised) Quality Assessment: In addition to purely algorithmic approaches, it is instrumental to work in collaboration with field scientists and experts, and incorporate in the quality assessment elements that experts indicate as important. Finally, there are a lot of exciting routes of collaboration with Human-Computer Interaction and Crowdsourcing experts, harnessing the "wisdom

of the crowds" in assessing the quality of our analysis, especially in applications such as *knowledge on the web* and *multi-aspect social networks*, where non-experts may be able to provide high quality judgements.

Application: Neurosemantics

In the search for understanding how semantic information is processed by the brain, I am planning to broaden the scope of the Neurosemantics applications, considering aspects such as *language*: are same concepts in different languages mapped in the same way in the brain? Are cultural idiosyncrasies reflected on the way that speakers of different languages represent information? Furthermore, I will consider more complex forms of stimuli (such as phrases, images, and video) and richer sources of semantic information, e.g, from Knowledge on the Web. There is profound scientific interest in answering the above research questions a fact also reflected on how well funded of a research area this is (e.g., see Brain Initiative `http://braininitiative.nih.gov/`)

Application: Urban & Social Computing

Social and physical interactionsof people in an urban environment, is an inherent multi-aspect process, that ties the physical and the on-line domains of human interaction. I plan to investigate human mobility patterns, e.g. through check-ins in on-line social networks, in combination with their social interactions and the content they create on-line, with specific emphasis on multi-lingual content which is becoming very prevalent due to population mobility. I also intend to develop anomaly detection which can point to fraud (e.g., people buying fake followers for their account). Improving user experience through identifying normal and anomalous patterns in human geo-social activity is an extremely important problem, both in terms of funding and research interest, as well as implications on revolutionizing modern societies.

Application: Knowledge on the Web

Knowledge bases are ubiquitous, providing taxonomies of human knowledge and facilitating web search. Many knowledge bases are extracted automatically, and as such they are noisy and incomplete. Furthermore, web content exists in multiple languages, which is inherently imbalanced and may result in imbalanced knowledge bases, consequently leading to skewed views of web knowledge per language. I plan to continue my work on web knowledge, devising and developing techniques that combine multiple, multilingual, structured (e.g., knowledge bases) and unstructured (e.g., plain text) sources of information on the web, aiming for high quality knowledge representation, as well as enrichment and curation of knowledge bases.

3) Funding, Collaborations, and Parting Thoughts

During my studies, I have partaken in grant proposal writing, contributing significantly to a successful NSF/NIH BIGDATA grant proposal (NSF IIS-1247489 & NIH 1R01GM108339-1) which resulted in $894,892 to CMU ($1.6M total) and has funded most of my PhD studies. In particular, I worked on outlining and describing the important algorithmic challenges as well as the applications we proposed to tackle. Being a major contributor to the proposal was a unique opportunity for me, giving me freedom to shape my research agenda. I have also been fortunate to have collaborated with a number of stellar researchers both in academia and industry, many of whom have been my mentors in research. Research agenda is very often shaped in wonderful ways through such collaborations, as well as through mentoring and advising graduate students. To that end, I intend to keep nurturing and strengthening my on-going collaborations, and pursue collaboration with scholars within and outside my field, always following my overarching theme, *bridging Signal Processing and Data Science for real-world applications*.

Selected References

[1] Rakesh Agrawal, Behzad Golshan, and **Evangelos E. Papalexakis**. A study of distinctiveness in web results of two search engines. In *WWW'15 Web Science Track*, (*author order is alphabetical*).

[2] Rakesh Agrawal, Behzad Golshan, and **Evangelos E. Papalexakis**. Whither social networks for web search? In *ACM KDD'15*, (*author order is alphabetical*).

[3] Miguel Araujo, Spiros Papadimitriou, Stephan Günnemann, Christos Faloutsos, Prithwish Basu, Ananthram Swami, **Evangelos E. Papalexakis**, and Danai Koutra. Com2: Fast automatic discovery of temporal ('comet') communities. In *Advances in Knowledge Discovery and Data Mining*. Springer, 2014.

[4] Alex Beutel, Abhimanu Kumar, **Evangelos E. Papalexakis**, Partha Pratim Talukdar, Christos Faloutsos, and Eric P Xing. Flexifact: Scalable flexible factorization of coupled tensors on hadoop. In *SIAM SDM'14*, 2014.

[5] **Evangelos E. Papalexakis**, Leman Akoglu, and Dino Ienco. Do more views of a graph help? community detection and clustering in multi-graphs. In *IEEE FUSION'13*.

[6] **Evangelos E. Papalexakis**, Alex Beutel, and Peter Steenkiste. Network anomaly detection using co-clustering. In *Encyclopedia of Social Network Analysis and Mining*. Springer, 2014.

[7] **Evangelos E. Papalexakis** and A. Seza Doğruöz. Understanding multilingual social networks in online immigrant communities. WWW '15 Companion.

[8] **Evangelos E. Papalexakis** and C. Faloutsos. Fast efficient and scalable core consistency diagnostic for the parafac decomposition for big sparse tensors. In *IEEE ICASSP'15*.

[9] **Evangelos E. Papalexakis**, Christos Faloutsos, and Nicholas D Sidiropoulos. Parcube: Sparse parallelizable tensor decompositions. In *ECML-PKDD'12*.

[10] **Evangelos E. Papalexakis**., Alona Fyshe, Nicholas D. Sidiropoulos, Partha Pratim Talukdar, Tom M. Mitchell, and Christos Faloutsos. Good-enough brain model: Challenges, algorithms and discoveries in multi-subject experiments. In *ACM KDD'14*.

[11] **Evangelos E. Papalexakis**, Tom M Mitchell, Nicholas D Sidiropoulos, Christos Faloutsos, Partha Pratim Talukdar, and Brian Murphy. Turbo-smt: Accelerating coupled sparse matrix-tensor factorizations by 200x. In *SIAM SDM'14*.

[12] **Evangelos E. Papalexakis**, Konstantinos Pelechrinis, and Christos Faloutsos. Location based social network analysis using tensors and signal processing tools. In *IEEE CAMSAP'15*.

[13] **Evangelos E. Papalexakis**, Nicholas D Sidiropoulos, and Rasmus Bro. From k-means to higher-way co-clustering: multilinear decomposition with sparse latent factors. *IEEE Transactions on Signal Processing*, 2013.

[14] Matt Gardner, Kejun Huang, **Evangelos E. Papalexakis**., Xiao Fu, Partha Talukdar, Christos Faloutsos, Nicholas Sidiropoulos, and Tom Mitchell. Translation invariant word embeddings. In *EMNLP'15*.

[15] U Kang, **Evangelos E. Papalexakis**, Abhay Harpale, and Christos Faloutsos. Gigatensor: scaling tensor analysis up by 100 times-algorithms and discoveries. In *ACM KDD'12*.

[16] Ching-Hao Mao, Chung-Jung Wu, **Evangelos E. Papalexakis**, Christos Faloutsos, and Tien-Cheu Kao. Malspot: Multi2 malicious network behavior patterns analysis. In *PAKDD'14*.

[17] N Sidiropoulos, **Evangelos E. Papalexakis**, and C Faloutsos. Parallel randomly compressed cubes: A scalable distributed architecture for big tensor decomposition. *IEEE Signal Processing Magazine*, 2014.