Revealing the collaborative dynamics of a large-scale arXiv text collection by means of k-shell decomposition

JAVIER VERA*, WriteWise, WriteWise Research Group, Artificial Intelligence Unit, Chile
WENCESLAO PALMA, Pontificia Universidad Católica de Valparaíso, Escuela de Ingeniería Informática, Chile
HECTOR ALLENDE, Pontificia Universidad Católica de Valparaíso, Escuela de Ingeniería Informática, Chile
SEBASTIAN RODRIGUEZ, Pontificia Universidad Católica de Valparaíso, Escuela de Ingeniería Informática, Chile
JUAN PAVEZ, Universidad Técnica Federico Santa María, Chile
EDUARDO FUENTES*, WriteWise, WriteWise Research Group, Artificial Intelligence Unit, Chile

1 INTRODUCTION

In recent years, a new paradigm has been gaining quickly interest within the scientific community: the *arXiv* repository. In this web platform, users share and classify their own preprints in a decentralized and open way. Users interact with this platform by sharing *TeX* files and are prompted to choose from a predefined set of tags representing areas of knowledge (for example, *computer science* or *statistics*). At the global level, a category evolves in time with no explicit central coordination from the microscopic activity of users sharing content. Moreover, this activity is open: all the shared content is freely available for users.

Regarding a specific category of *arXiv*, the process of content sharing can be viewed as a *semiotic dynamics* phenomena, in which populations of users develop a communication system (see, for example, [5]). Indeed, the accumulation of preprints over time is a way to densify and expand the meaning of the particular category.

An extended framework to reveal the development of complex decentralized patterns of (human) behavior is *network theory*. Particularly, previous work on language networks has revealed *small world* properties (see, for example, [4; 6]). These studies, however, only focused on the individual degree or centrality of a given node, not taking into account the relationships with other nodes.

To overcome these problems, the analysis known as $k-shell$ decomposition highlights structural properties of (language) networks not captured by classical graph measures. By removing sets of nodes of increasing degree, the $k-shell$ decomposition process [1; 2; 8] identifies layers of nodes of similar connectivity, called $k-shells$, each one obtained by a recursive pruning strategy that reveals the hierarchical organization of the network.

In this work, we used the $k-core$ decomposition analysis for the study of language networks representing activity patterns of *arXiv* users, within the *computer science* category. The main goal is to understand the evolution of a scientific community arising from the microscopic activity of users posting content.

Authors' addresses: jv@biopub.cl; Javier Vera*, WriteWise, WriteWise Research Group, Artificial Intelligence Unit, Av. Beaucheff 935, Santiago, Chile; Wenceslao Palma, Pontificia Universidad Católica de Valparaíso, Escuela de Ingeniería Informática, Avenida Brasil 2241, Valparaíso, Chile; Hector Allende, Pontificia Universidad Católica de Valparaíso, Escuela de Ingeniería Informática, Av. Brasil 2241, Valparaíso, Chile; Sebastian Rodriguez, Pontificia Universidad Católica de Valparaíso, Escuela de Ingeniería Informática, Av. Brasil 2241, Valparaíso, Chile; Juan Pavez, Universidad Técnica Federico Santa María, Av. España 1680, Valparaíso, Chile, ef@writewise.cl; Eduardo Fuentes*, WriteWise, WriteWise Research Group, Artificial Intelligence Unit, Santiago, Chile.

2 NOTATIONS AND DEFINITIONS

$k-shell$ decomposition

Let us consider an undirected and unweighted *graph* $G = (V, E)$, where V denotes the set of $|V| = n$ *nodes* and E denotes the set of $|E| = e$ *edges*. Now, we introduce some technical definitions from [1; 2].

Definition 2.1. A subgraph $H \subseteq G$ is a $k-core$ C_k of G if and only if H is the largest subgraph whose nodes have a degree of at least k. The $k-core$ of G may be obtained by removing all the nodes from G of degree less than k, until all remaining nodes have degree at least k.

Definition 2.2. A node $u \in V$ has *shell index* k if it belongs to the $k-core$ but not to the $(k+1)-core$. The $k-shell$ S_k is formed by all nodes whose *shell index* is k. The $k-core$ is therefore the union of all shells S_r, such that, $r \geq k$, that is,

$$C_k = \bigcup_{r \geq k} S_r$$

.

The $k-core$ decomposition therefore decomposes the graph in successive layers, revealing groups of nodes with their own density patterns from the outmost one to the most internal one (the $S_{k_{max}}$).

Graph-of-words (GoW)

We adopted a simple version of the Graph-of-Words representation (GoW) of texts [3; 7]. A text is seen as a undirected and unweighted graph where nodes are unique nouns, and where there is an edge between two nodes if they co-occurr within a sentence. In our case, we transformed a collection of texts C in a set of sentences $S = \{s_i\}_i$. In the first place, we identified the set of unique nouns $W = \cup_{s \in S} s$. Secondly, through an iterative process we inspected all sentences of S in order to find co-occurrences between pairs of nodes.

3 RESULTS

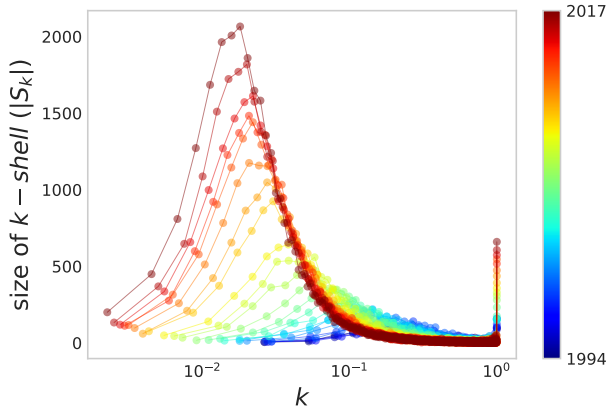
The dataset used in this study was obtained using a web scraper[1] for specific data ranges and categories from the electronic repository of preprints *arXiv*[2]. Specifically, abstracts from the *computer science* category from 1994 until 2017 were used.

The activity of users interacting with the *arXiv* system consists to add new preprints to the system. To add a new preprint, besides

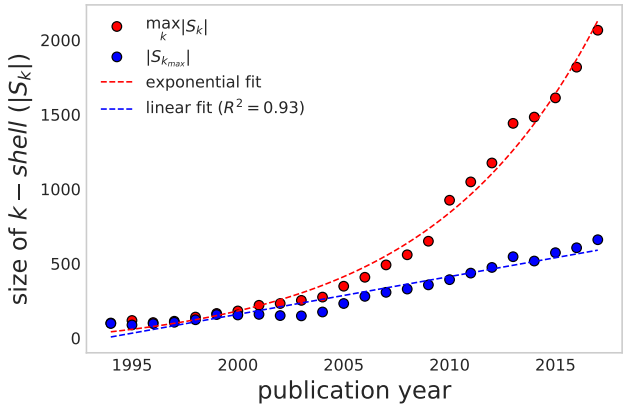
[1] https://github.com/Mahdisadjadi/arxivscraper
[2] https://arxiv.org/

some *metadata* (authors, title, among others) the user is prompted for a set of tags indicating knowledge categories[3].

For text preprocessing (sentence boundary detection, stopword removal, noun filtering and lemmatization), we used the libraries *NLTK*[4] and *spaCy*[5]. Graph construction and $k-core$ decomposition analysis was made using *NetworkX*[6] and *NetworKit*[7].



(a) $|S_k|$ over time. For each year, it is showed the size of the $k-shell$ ($|S_k|$) versus the (normalized) index k.



(b) $|S_{k\,max}|$ and $\max_k |S_k|$ over time. For each year, it is showed the evolution of $|S_{k\,max}|$ and the size of $|S_k|$ at the critical value $k^* = \arg\max_k |S_k|$.

Fig. 1. Analysis of the $k-core$ decomposition over the time interval 1994-2017.

The first observation of the structure of the $k-core$ decomposition is shown in Figure 1 (a). This figure shows the size of of $k-shell$ S_k as a function of its normalized index k, for each graph representing the collection of texts of each year from 1994 to 2017. For each $k-core$ decomposition, the behavior of $|S_k|$ versus k exhibits three clear domains. First, $|S_k|$ reached a maximum at $k = k^*$. This critical value $k*$ strongly depended on the year. Indeed, k^* tended to exhibits a negative exponential trend over time. Second, a drastic decreasing of $|S_k|$ is observed for $k > k^*$, until a stationary value ~ 0. Finally, a drastic increasing in $|S_k|$ was found at $k = k_{max}$.

The properties of the successive $k-shells$ for the graphs constructed for each year can be studied by describing $|S_k|$ over time. Figure 1 (b) shows the evolution of two interrelated quantities: the size of S_{k^*} (where k^* is the critical value of k^*) and the size of $S_{k_{max}}$. Remarkably, $S_{k_{max}}$ followed a slow positive linear trend (with slope ~ 0.04), whereas $|S_{k^*}|$ followed a exponential law $|S_k^*| \sim year^{1.24}$.

4 CONCLUSIONS

In this work was shown how $k-shell$ decomposition helps to understand the dynamics of the formation of the decentralized and collaborative language community defined by the electronic repository *arXiv*. Our results suggest that there are several global patterns that emerges from the microscopic activity of users sharing content. The growth of the collection of texts (and therefore of the associated networks) was (almost) completely governed by the outmost $k-shells$, which exponentially increased its size over time. Nevertheless, the size of the most dense set of nodes ($S_{k_{max}}$) tends to linearly increase its size. This points in the direction of the existence of an exponential accumulation of words that forces changes in the main discipline (*computer science*, in our case), represented by $S_{k_{max}}$. These observations were confirmed by the behavior of the (normalized) critical index $k^* = \arg\max_k |S_k|$, since it exponentially shifts to the outmost network layers. Further study should describe the relationship between the index k and the number of connected components of the $k-shell$ S_k. Moreover, it is plausible to propose that the decentralized features of *arXiv* appear precisely at those external layers.

ACKNOWLEDGMENTS

This work was supported by Innova Corfo Chile: Capital Humano para la Innovación, Grant CH17-83813 (to J. Vera and E. Fuentes).

REFERENCES

[1] José Alvarez-Hamelin, Luca Dall'Asta, Alain Barrat, and Alessandro Vespignani. K-core decomposition of internet graphs: hierarchies, self-similarity and measurement biases. *Networks and Heterogeneous Media*, 3(2):371–393, 2008. Exported from https://app.dimensions.ai on 2018/10/08.

[2] Vladimir Batagelj and Matjaz Zaversnik. Generalized cores. *CoRR*, cs.DS/0202039, 2002.

[3] Roi Blanco and Christina Lioma. Graph-based term weighting for information retrieval. *Inf. Retr.*, 15(1):54–92, February 2012.

[4] Ramon Ferrer i Cancho and Richard V. Solé. The small world of human language. *Proceedings of the Royal Society of London B: Biological Sciences*, 268(1482):2261–2265, 2001.

[5] Ciro Cattuto, Vittorio Loreto, and Luciano Pietronero. Semiotic dynamics and collaborative tagging. *Proceedings of the National Academy of Sciences*, 104(5):1461–1464, 2007.

[6] Ramon Ferrer i Cancho, Ricard V. Solé, and Reinhard Köhler. Patterns in syntactic dependency networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 69 5 Pt 1:051915, 2004.

[7] R. Mihalcea and P. Tarau. TextRank: Bringing order into texts. In *Proceedings of EMNLP-04and the 2004 Conference on Empirical Methods in Natural Language Processing*, July 2004.

[8] Stephen B. Seidman. Network structure and minimum degree. *Social Networks*, 5(3):269 – 287, 1983.

[3] see http://arxitics.com/help/categories for a detailed list
[4] https://www.nltk.org/
[5] https://spacy.io/
[6] https://networkx.github.io/
[7] https://networkit.iti.kit.edu/