A Network model of the Chemical Space provides similarity structure to the system of chemical elements

Eugenio Llanos[1,2,3], Wilmer Leal[1,2] Andrés Bernal[2,4], Guillermo Restrepo[2], Jürgen Jost[2], and Peter F. Stadler[1,2,5]

[1] Bioinformatics Group, Department of Computer Science, Universität Leipzig,Härtelstraße 16-18, 04107 Leipzig, Germany
ellanos@sciocorp.org,
[2] Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, 04103 Leipzig, Germany
[3] Corporacióon SCIO, Calle 57b 50-50 bloque d22 of. 412, 111321 Bogota, Colombia
[4] Department of Basic Sciences, Universidad Jorge Tadeo Lozano, 110311 Bogota, Colombia
[5] The Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, New Mexico 87501

1 Introduction

The collection of every species reported up to date constitutes the so-called Chemical Space (CS). This space currently comprises well over 30 million substances and is growing exponentially [2]. In order to characterize this ever-growing space, chemists seek for similarity of substances on the CS based on the way they combine [3]. Mendeleev's work on chemical elements was based upon his knowledge of the CS by 1869 is perhaps the most famous example of how the CS determines similarity relations [4]. From a contemporary point of view, Network Theory serves as a natural framework to identify these kind of relational patterns in the CS [5]. Nowadays, databases such as Reaxys©[6] have grown to a point where they can be taken as proxies for the whole CS, opening the possibility to analyze it from a data driven perspective.

In this work we propose to study the similarity of chemical elements according to the compounds they form. From each compound, we deleted each element to obtain a formula that is connected to the deleted element, v.g. $S_{1/2}O_{4/2}$, $Na_{2/1}O_{4/1}$ and $Na_{2/4}S_{1/4}$ are formulae coming from Na_2SO_4 (Sodium sulfate) where Na, S and O, have been deleted respectively. This form a bipartite graph formed by elements and those formulae where they have been deleted, We build our network using 26,206,663 compounds recorded on Reaxys up to 2015. Similarity among chemical elements is constructed analogously to Social Network Analysis, where actors are declared similar whenever they are connected to the same set of other actors. The more formulae elements share, the more similar they are. We introduce a new notion of in-betweenness of elements acting as *mediators* on similarity relations of others. We analyze the structural features of this network and how they are affected by node removal. We show that the network is both highly dense and redundant. Even though it is heavily centralized,

similarity relations are widely spread across a wide range of formulae, which grants the network extraordinary structure resiliency, even against directed attack. We discuss some implications of these results for chemistry.

2 Results

- *The network is heavily centralized*: chemical reactivity of elements is far from uniform, as the degree distribution of elements exhibits three different regions, see Figure 1(b). The first one is composed by a few elements that concentrate the vast majority of relations (the first is H, which accounts for 95.9% of formulae, followed by C 95%.0). The second one is composed by the bulk of elements, which connect to 10,000-100,000 formulae. The third region corresponds to elements that have a very low number of molecular formulae.
- *Formulae with degree one are mostly connected to central elements*: formulae of degree one correspond to compounds that are unique to one element (*singularities*). Eight elements concentrate most singularities (90%) evoke both the singularity principle of the periodic chart and the distinction between organic and inorganic chemistry. In general, the number of singularities scales semi-linearly with element degree (power law with exponent 1.3, see Figure 1(a)). This result shows that elements tend to be unique as long as more compounds of them are obtained, independently of their identity. The more compounds one element has, the less similar to others it becomes.
- *Similarity does not partition the space into clear-cut classes of elements*: since formulae generate similarity relations among the elements that are connected to them, the degree of one formula corresponds to the number of elements it makes similar. The smoothness of this degree distribution (Figure 1(a)) shows that elements cannot be divided into clear-cut classes, since otherwise such classes would produce local maxima corresponding to the sizes of these classes. This result has an interesting chemical implication, as it challenges the usual view of elements as separated *families*.
- *Element in-betweenness depends on its degree*: elements work as mediators of similarity relations through the formulae they constitute. Such mediation scales almost linearly with the degree of the element (see Figure 1(c)). This is a very interesting feature, since it shows that similarity relations are not concentrated on certain kind of compounds or manifested by specific elements working as mediators, but are evident on the entire CS.
- *Similarity relations are highly resilient to directed attack*: since the network is highly centralized, deleting random elements should not have a major effect on the network topology. We instead deleted sequentially elements from the one with highest degree down to 12 elements and those formulae on which they take part. Deleting central elements has impact on the degree of the elements and the distribution goes down on absolute frequency. Notwithstanding, almost all elements are affected in the same way and the shape of the curve is conserved (see different data series on Figure 1(b)). The same happens on the degree of formulae, which is shifted towards the left, but the shape remains (Figure 1(a)).

– *Strong and weak similarity relations are the less variant*: since our network is of an epistemic nature, vulnerability can be related to the viability of extracting knowledge with limited information. To test how variant are the similarity relations against removal of molecular formulae, we calculated the variance of the rank of pairwise element similarity (number of length 2 paths between the corresponding nodes) when keeping only similarities mediated by each element. Surprisingly, strong and weak similarities have the lowest variance (see Figure 1(d)), showing that similarities are by no means random but they form a strong structure that stands across the entire CS, revealing a fundamental nature of these similarity patterns.

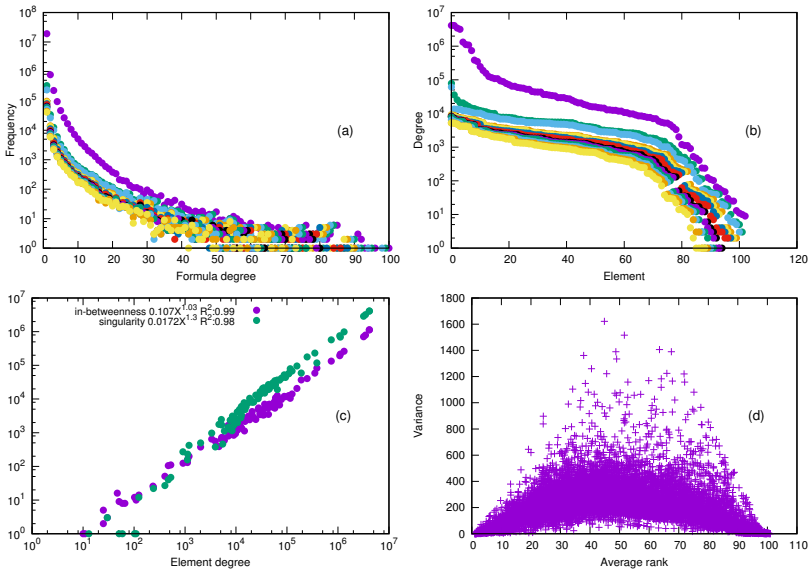

Fig. 1. (a) Distribution of formula degree. Different colors of points correspond to series where different central elements has been removed. (b) Degrees of elements. (c) Singularities and in-betweennes vs formula degree. (d) Variance of pairwaise rank position vs average rank position. Low variance is found on low average rank positions (similar elements) and high average rank positions (dissimilar elements).

References

1. Schummer, J.: Scientometric studies on chemistry II: Aims and methods of producing new chemical substances. Scientometrics 39 (1), 125–140 (1997)
2. Llanos, E.; Leal, W.; Luu, D.; Jost, J.; Stadler, P.F.; Restrepo, G.: Exploration of the chemical space and its three historical regimes. Proceedings of the National Academy of Sciences of the United States of America 116 (26), 12660–12665 (2019) https://doi.org/10.1073/pnas.1816039116

3. Schummer, J.: The chemical core of chemistry I: a conceptual approach. HYLE–International Journal for Philosophy of Chemistry 4 (2), 129-162 (1998)
4. Leal, W.; Llanos, E.; Stadler, P.F.; Jost, J.; Restrepo, G.: The Chemical Space from Which the Periodic System Arose. ChemRxiv 10.26434/chemrxiv.9698888.v1 (2019)
5. Leal, W.; Restrepo, G.; Bernal, A.: A network study of chemical elements: from binary compounds to chemical trends. MATCH communications in mathematical and in computer chemistry 68, 417–442 (2012)