Quantifying disruptiveness using a neural embedding method

Keywords: Graph embedding, science of science, disruptiveness, innovation, bibliometrics

Extended Abstract

Development and *disruption* are a common dichotomy for science and technology; Scientific papers or products may consolidate and improve existing works (development), or open up a new field or redefine a product (disruption). Given the outsized impact of disruptive creations and the reported decrease in disruptive science [1], understanding the mechanisms and conditions that spur disruptive innovations (e.g., the role of small teams studied by Wu et al. [2] or the impact of scientific fields investigated by Chu et al. [3]) has been attracting much attention.

Yet, although the disruption index (denoted as *D*) by Funk and Owen-Smith [4] played a pivotal role in the investigation of disruptive innovations, the characteristics of the disruptiveness measure are not fully understood, and limitations of the measure have been identified. For instance, Bornmann et al. found that the disruption index of the papers in bibliometrics is concentrated around zero and only a small fraction of papers shows significant variance [5]. A related issue is that the index is restricted to the immediate neighbors of a focal paper and thus overlooks any topological structures that its larger context may exhibit [6]. The disruption index's reliance on the immediate local structure—often determined by few papers—raises a question: can we define a better alternative to the disruption index that overcomes the major limitations of the index?

Here, we introduce a graph embedding approach to quantify disruptiveness and demonstrate that our measure outperforms the original disruption index in multiple tasks. Our measure shares the same idea with the disruption index—because disruptive papers create a new knowledge space, the papers that cite the disruptive papers are much less likely to cite the references of the disruptive papers. We leverage the fact that the *target* and *context* vectors that node2vec (word2vec) learns have different semantics: by restricting the training context to only one side of the target node, we can learn *the context of citations* and *that of references* separately. In other words, in a directed node2vec, the target vector of a paper would capture the context of the citations of the paper while the context vector of the paper would capture the context of the references (Fig. 1a). This model estimates the probability that paper i is cited by paper j

$$P(j|i) = \frac{\exp(u_j \cdot v_i)}{Z_i}$$

, where v_i is the target vector of paper i and u_j is the context vector of the paper j. Therefore, the target vector of a paper is trained to have high cosine similarity with the context vectors of its *citations* and the context vector of a paper is trained to have high cosine similarity with the target vectors of its *references*. In addition to this, when a paper is less disruptive, many of its citations also cite its references and this makes the target vectors of the references similar to the context vectors of citations (Fig. 1b).

Taken all together, the cosine distance between the target vector and context vector of less disruptive papers tends to be smaller than disruptive papers citations. Therefore, we define the new disruption index of the paper i as

$$CD_i = 1 - \frac{u_i \cdot v_i}{|u_i||v_i|}$$

which is the cosine distance between the target vector of paper i (v_i) and the context vector of paper i (u_i).

We applied our measure to the citation network in the APS dataset consisting of 644,022 papers published from 1900 to 2019 and 8,323,911 citation connections. Fig. 1c shows the distribution of our disruptive index CD and the previous index by Funk and Owen-Smith D for all papers in our dataset. We observe that, in contrast to the D measure concentrated on zeros, the new measure CD shows a higher resolution around zeros.

Then we compared the new measure CD to the traditional measure D, in identifying disruptive papers such as Nobel Prize and APS milestone papers, as well as disruptive papers selected through a scholar survey. Our proposed measure placed the Nobel prize papers near the top 20 % while the disruption index put them in both the top and bottom 10 % (Fig. 1d). To ensure that these results were not solely due to the number of citations and references, both measures are calculated again based on a random network while preserving the number of citations and references. The results showed that CD values from the random network were completely different from that of the real network while the distribution of D values based on the random network remained concentrated in the top and bottom 10 %. The same pattern is observed for the APS's "milestone papers" (Fig. 1e). Next, we examined the disruptive papers selected through the scholar survey [2]. The CD values of these papers were top 10 % and top 21 %, while D values of those papers lie in the bottom 0.2 % and top 42 %, showing that CD outperforms D. Furthermore, CD values of papers that headline prominent prior works, which are expected to be less disruptive [2], were significantly smaller when they are calculated based on the real network compared to random networks (Fig. 1f, $p \ll 10^{-3}$, Cohen's D = -0.228). In contrast, D values of these articles based on the real network were found to be significantly larger than those obtained from random networks, although the effect size was small (Fig. 1f, $p \ll 10^{-3}$, Cohen's D = 0.026). The better identification of disruptive papers and higher resolution of our new disruption index will provide more insights into the studies of disruptive innovations.

References

[1] M. a. L. E. a. F. R. J. Park, "Papers and patents are becoming less disruptive over time," *Nature,* vol. 613, no. 7942, pp. 138--144, 2023.

[2] L. a. W. D. a. E. J. A. Wu, "Large teams develop and small teams disrupt science and technology," *Nature,* vol. 566, no. 7744, pp. 378--382, 2019.

[3] J. S. a. E. J. A. Chu, "Slowed canonical progress in large fields of science," *Proceedings of the National Academy of Sciences,* vol. 118, no. 41, p. e2021636118, 2021.

[4] R. J. a. O.-S. J. Funk, "A dynamic network measure of technological change," *Management science,* vol. 63, no. 3, pp. 791--817, 2017.

[5] L. a. T. A. Bornmann, "Disruptive papers published in Scientometrics," *Scientometrics,* vol. 120, no. 1, pp. 331--336, 2019.

[6] C. a. C. Q. a. Y. E. a. B. Y. a. S. J. Min, "Citation cascade and the evolution of topic relevance," *Journal of the Association for Information Science and Technology,* vol. 72, no. 1, pp. 110--127, 2021.

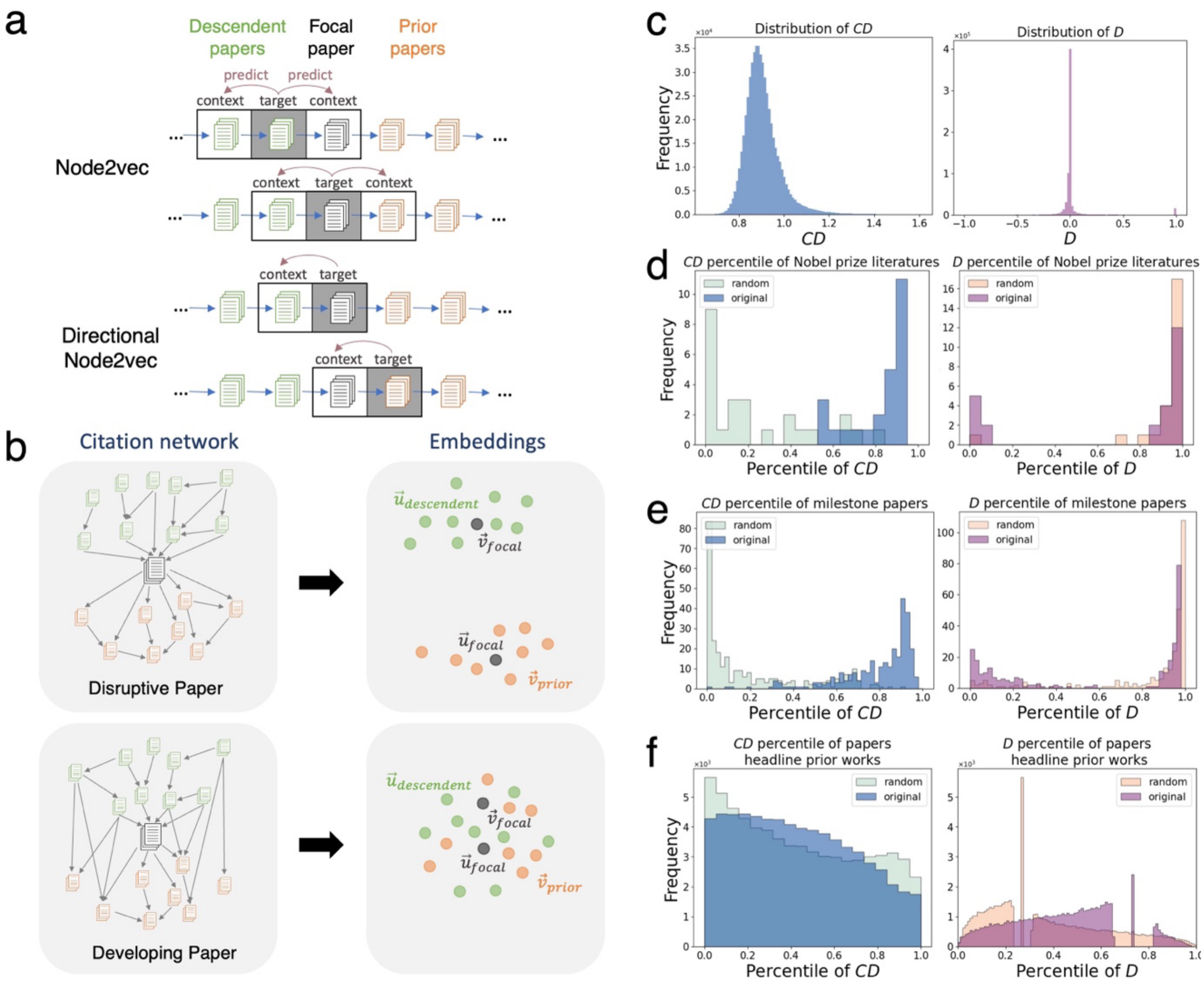

Figure 1: New disruption index using graph embedding. (a) Node2vec and Directional Node2vec generate random walks following the citation relationship (blue arrows) on the citation network. Based on the random walk, both models estimate the probability of predicting context papers given the target paper. The papers surrounding a target paper in both directions are context papers in Node2vec while the context papers in Directional Node2vec exist in only one direction. (b) The cosine distance between the target vector and context vector is closer when a paper is less disruptive. (c-f) The distribution of a new disruption index *CD* and a previous index *D* of the whole data set papers, 25 Nobel prize papers, 283 milestone papers, and 67,292 papers headline prior works calculated from the real citation network and the random citation network.