Predicting Emerging Technologies: A Text Mining Approach Based on The Temporal Exponential Random Graph Model (TERGM) Methodology.

Predicting emerging technologies is useful for research foundations and policy makers aiming to promote and enhance the development of potentially promising research fields. However, the current approaches based on scientometrics often lack a robust predictive ability because most have been limited to ex post evaluation which measures past performance or impacts, and which are vulnerable to bear the risk of "survivorship bias".

Rotolo (2015) reviewed some quantitative studies related to emerging technologies, and further summarized the five attributes of emerging technologies: radical novelty; relatively fast growth; coherence; prominent impact; uncertainty and ambiguity [1]. Though these five factors are already considerate, but obviously, self-evolution process of emerging technologies are ignored. In fact, technological evolution can be conceived as a recombinatory process of novel and existing component technologies within complex adaptive systems [2]. Thus, emerging technologies also can be conceived as a process of formation of technological recombination via a network evolution perspective.

As a remedy, we propose a text mining approach based on the temporal exponential random graph model to identify emerging technologies on the contest dataset of 9th Global Tech-Mining Conference. The temporal exponential random graph model, usually called TERGM, is an extension of the ERGM designed to accommodate inter-temporal dependence in longitudinally observed networks [3]. The essential differences from the previous studies are focusing more on formation of technological recombination, which makes it possible to use network inference method to estimate the odds of node pair formation, herein referred to the MeSH-pairs on the time series scenario. And moreover, the growth ratio of the MeSH-pairs, rather than of technological terms, compared to the previous time period is used to determine technology's emergence.

The essential part of modeling TERGM as follows: (1) TERGM assumes that the statistics formed based on the previous stage of networks fully encompass the dependencies observed in the network at current stage. So, the whole dataset were sliced into five snapshots by a 2 year time interval. (2) Building MeSH term co-occurrence network and reducing the size of network for each time slice. (3) Observed network selection is a precondition for a good simulation-based model, which influences the theoretical accuracy of model. Here, the hypothesis was proposed to constrain terms emerging conditions. Those MeSH-pairs have to remain 50% growth ratio compared to the previous slice can be conceived as emergence. (4) Predictors of models mainly contain the structural features of network, such as the numbers of edges, triangle, degree centrality and betweenness centrality also be considered. It is worth mentioning at this point that stability across slices and MeSH similarity based on documents were accounted for. (5) Statistical inference is the important step for TERGM building. Established models can be checked by comparing simulated networks based on the model with the actual observed networks. Here, the out-of-sample goodness-of-fit assessment is applied to TERGM, that is, we estimated the model based on the first four networks, and predict the last network that were left out. The resulting diagram is shown in Figure 1. The first three subfigures compare the distribution of observed and simulated endogenous network statistics. The last subfigure presents receiver operating characteristics (ROC) and precision–recall (PR) curves. The PR curve shows that the predictive fit more than 90%. (6) Based on the above assessment, the whole dataset that contains five slices was applied to train a TERGM, thus the emerging probability of MeSH-pairs at the next stage was predicted.

Finally, we trained an embedding model to represent relationship between MeSH-pairs and technological terms, and those terms have the closest relationship with emerging MeSH-pairs are chosen as the best representative of the emerging technology terms [4].

References:

[1] ROTOLO D, HICKS D, MARTIN B R. What is an emerging technology? [J]. Research Policy, 2015, 44(10): 1827–1843.
[2] FLEMING L, SORENSON O. Technology as a complex adaptive system: evidence from patent data[J]. Research Policy, 2001, 30(7): 1019–1039.
[3] LEIFELD P, CRANMER S J, DESMARAIS B A. Temporal Exponential Random Graph Models with btergm: Estimation and Bootstrap Confidence Intervals [J]. Journal of Statistical Software, 2017, VV(2): 1–36.
[4] WU L, FISCH A, CHOPRA S. StarSpace: Embed All The Things! [C]. The Thirty-Second AAAI Conference on Artificial Intelligence//2018: 1–9.

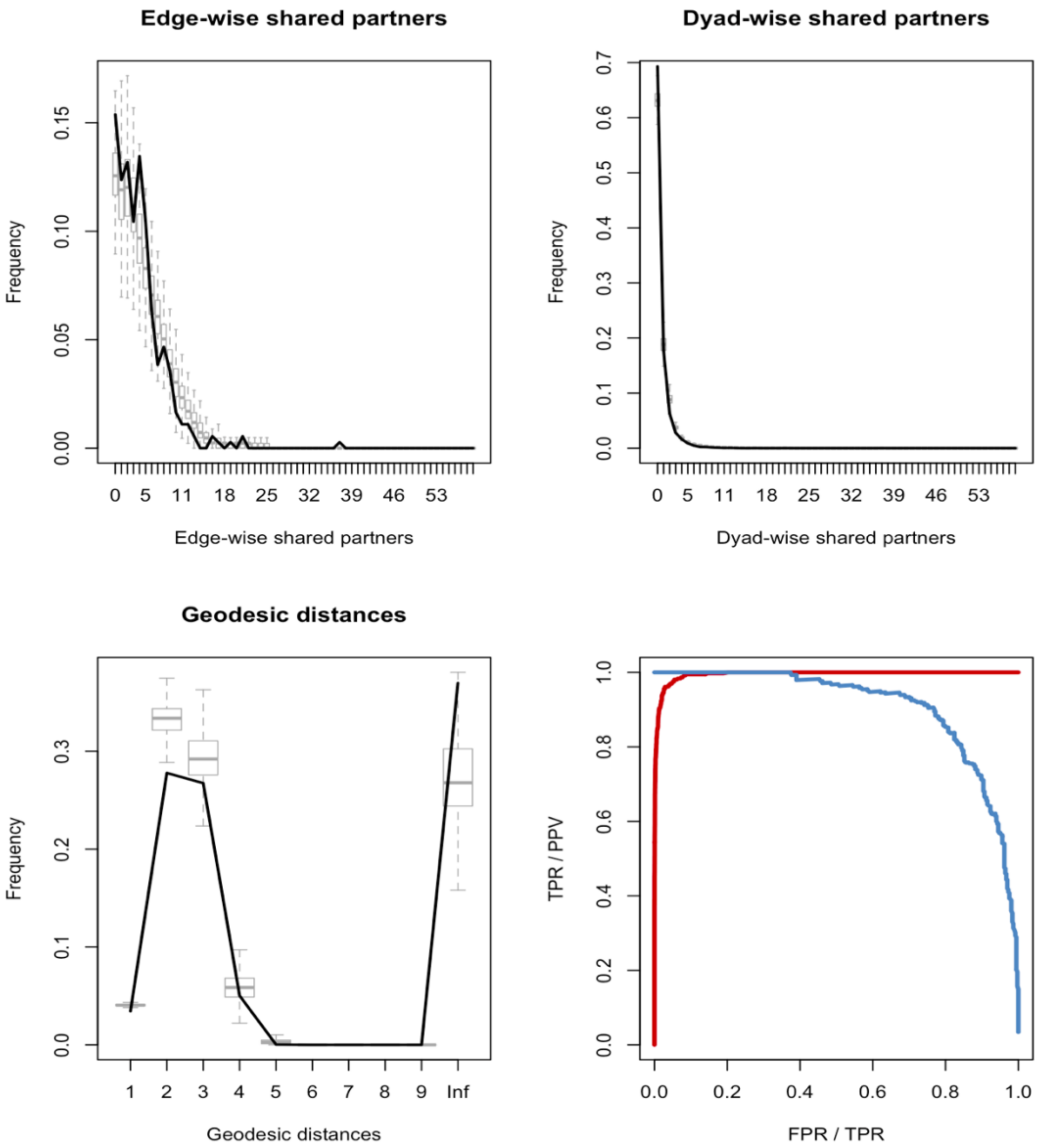

Figure 1: Out-of-sample goodness-of-fit assessment for TERGM.