PLOS COMPUTATIONAL BIOLOGY

PERSPECTIVE

Network motifs and their origins

Lewi Stone[1,2☯]*, **Daniel Simberloff**[3☯], **Yael Artzy-Randrup**[4,5☯]

1 Biomathematics Unit, School of Zoology, Faculty of Life Sciences, Tel Aviv University, Israel, **2** Mathematical Sciences, School of Science, RMIT University, Melbourne, Australia, **3** Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, Tennessee, United States of America, **4** Department of Theoretical and Computational Ecology, IBED, University of Amsterdam, Amsterdam, the Netherlands, **5** Institute of Advanced Study, University of Amsterdam, Amsterdam, the Netherlands

☯ These authors contributed equally to this work.
* lewistone100@gmail.com

Author summary

Modern network science is a new and exciting research field that has transformed the study of complex systems over the last 2 decades. Of particular interest is the identification of small "network motifs" that might be embedded in a larger network and that indicate the presence of evolutionary design principles or have an overly influential role on system-wide dynamics. Motifs are patterns of interconnections, or subgraphs, that appear in an observed network significantly more often than in compatible randomized networks. The concept of network motifs was introduced into Systems Biology by Milo, Alon and colleagues in 2002, quickly revolutionized the field, and it has had a huge impact in wider scientific domains ever since. Here, we argue that the same concept and tools for the detection of motifs were well known in the ecological literature decades into the last century, a fact that is generally not recognized. We review the early history of network motifs, their evolution in the mathematics literature, and their recent rediscoveries.

Citation: Stone L, Simberloff D, Artzy-Randrup Y (2019) Network motifs and their origins. PLoS Comput Biol 15(4): e1006749. https://doi.org/10.1371/journal.pcbi.1006749

Editor: Ruth Nussinov, National Cancer Institute, United States of America and Tel Aviv University, Israel, UNITED STATES

Published: April 11, 2019

Funding: The support of the Australian Research Commission (https://www.arc.gov.au/) for grants DP15102472 and DP170102303 is gratefully acknowledged. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript

Competing interests: The authors have declared that no competing interests exist

Complex networks now feature prominently in many aspects of modern science and society [1–18]. Within this still rapidly growing discipline, there is strong recognition that large-scale dynamical properties of a network are governed by its much smaller constituent "network motifs," and so the chief focus is often on motifs. In more technical terms, motifs may be defined as "patterns of interconnections (or subgraphs) occurring in complex networks at numbers significantly higher than those in randomized networks" [1]. Their presence indicates the operation of underlying nonrandom structural or evolutionary design principles that might have been involved in building the network. In a key paper by Milo and colleagues (2002) [1] entitled "Network motifs: Simple building blocks of complex networks," the authors propose a powerful technique for identifying nonrandom motifs that might otherwise remain hidden. The ability to detect network motifs has had far-reaching scientific impact, and the article of Milo, Alon and colleagues [1] and its three associated papers [2–4] are considered transformative in the field, as can be gauged from the approximately 860 citations they receive annually. However, it is not well known that the very same method used by Milo, Alon and colleagues [1] has a long history in the ecological and social sciences, as discussed here.

In practice, motifs arise in different contexts with diverse forms, as seen in Fig 1 showing a checkerboard motif used in the study of ecological networks [5] (and discussed in more detail below), a clustering triangular motif used in sociological and epidemiological contexts [6–9],

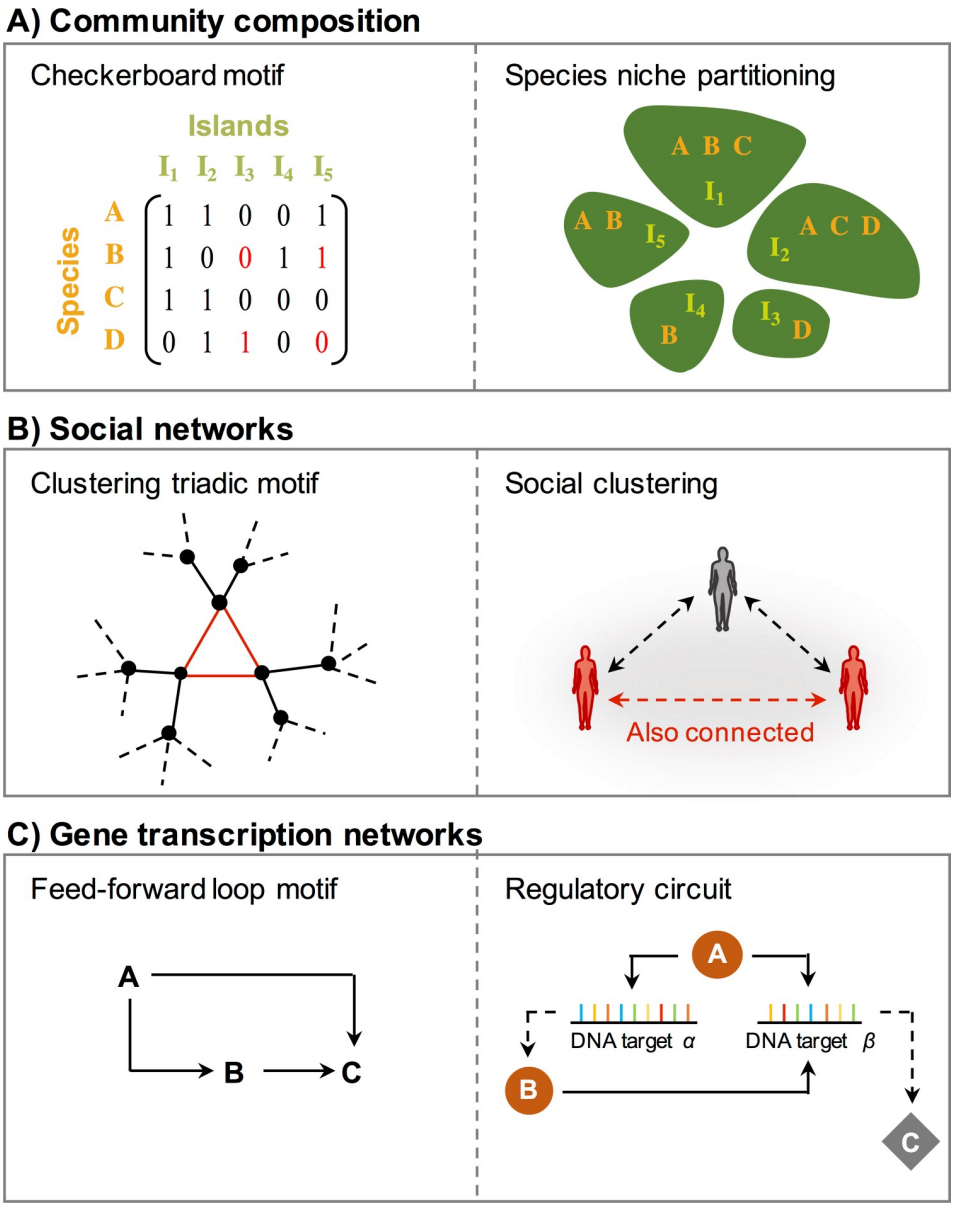

Fig 1. Network motif examples. Motifs in different contexts (right column) and example systems (left column). (A) Checkerboard motif. For example, 4 species (A–D) occupy 5 islands (I_1–I_5). The checkerboard motif highlighted in red represents 2 species that do not co-occur on the same island (here, B appears on I_5 but D does not, and conversely, D appears on I_3 but B does not), suggestive of competitive interactions. (B) Triadic clustering motif. For example, the motif represents cases in which an individual's connected friends are also connected with each other, having significance, for example, in social networks and epidemiological contact networks. (C) Feed-forward loop motif. For example, a circuit in gene transcription networks, in which DNA target β can be activated only through simultaneous binding of two transcription factors A and B, and in which B depends on A initially binding to DNA targets α and β, suggesting regulatory control on transcription.

https://doi.org/10.1371/journal.pcbi.1006749.g001

and a feed-forward loop motif commonly used in systems biology [1,3,10]. These are only a few examples. Algorithms that detect overabundant motifs have had many applications in systems biology in which they have been used, for example, in the search for regulatory algorithms [4] of the cell and for applications concerning cancer diagnosis [11]. The same techniques are also being applied to study neuronal networks [12], brain function [13], social

networks [14], financial [15] and trade networks [16], and internet and mobile wireless communication [17]. This has led to a whole range (quite likely hundreds) of software toolboxes or one-off programs for detecting network motifs [18]. This paper outlines briefly the origins and history of network motifs and the main algorithm for identifying motifs, that is, before its recent "rediscovery" [19] (Merton 1961) at the turn of the millennia by Milo and colleagues (2002) [1] and Shen-Orr and colleagues (2002) [3].

In order to proceed, we define a few basic terms from network theory. Any network or graph may be studied in terms of its binary adjacency matrix A. For a network with n nodes, and an $n \times n$ binary adjacency matrix A, then $A_{ij} = 1$ implies that node-i is connected to node-j, and $A_{ij} = 0$ otherwise. The network is undirected if $A_{ij} = A_{ji}$, but in general, we deal with directed networks in which this equality usually doesn't hold. The row and column sums of A are given by $r_i = \sum_{j=1}^{n} A_{ij}$, and $c_j = \sum_{i=1}^{n} A_{ij}$, and represent the in- and out-degrees of all nodes in the network. A key goal is to find a way to generate independent random samples from the full "universe" of all possible binary adjacency matrices that have the same row and column sums $r = (r_i)$, $c = (c_j)$, respectively. This universe of matrices is referred to as $U(r,c)$ and constitutes the universe of all possible matrices having the same row and column constraints, thus preserving an important topological feature of the observed adjacency matrix A.

In ecological applications, let us suppose for a given adjacency matrix A that rows represent species and columns represents islands. Then $A_{ij} = 1$ implies that species-i inhabits island-j. The random matrix ensemble should lock-in characteristics that reflect the ability of some species to colonize islands better than other species as well as the feature that some islands hold more species than others. For this reason, we generate a reference ensemble of random matrices for our null-model in which the row sums (reflecting species colonization abilities) and column sums (reflecting island species numbers) never change [20].

To our knowledge, the first rigorous methods for detecting nonrandom patterns in adjacency matrices or networks, when compared to the universe $U(r,c)$ of all possible matrices, can be traced to the works of Connor and Simberloff (1979) [20], Stone (1988) [21], and Stone and Roberts (1992) [5]. These studies use the so-called switch method to generate an ensemble of random matrices. The method randomly switches or interchanges checkerboard configurations, as shown in Fig 2, and rests on the observation that applying a single such switch leaves the row and column sums of the matrix unchanged. Applying enough switches randomizes the adjacency matrix, but with each switch the row and column sums of the matrix remain preserved. The latter are generally fixed to the values of the observed matrix being tested. Simberloff (1986) [22] and Stone (1988) [21] attempted to show computationally that the switch method randomly samples the universe of all possible matrices from $U(r,c)$ in a manner that is approximately uniform and discussed schemes for drawing samples after every k successive interchanges. Zaman and Simberloff (2002) [23] and Artzy-Randrup and Stone (2005) [24] show rigorously, using different methods, that exact uniformity can be achieved with easily implemented weighting schemes. Independent matrix samples so generated from the ensemble $U(r,c)$ provide a reference frame that can be used to estimate motif frequencies that should be expected with a random model. Other novel sampling methods have since been devised [25,26].

A method specifically for finding an over-represented "network motif" was, to our knowledge, first outlined explicitly in Stone (1988) [21] and Stone and Roberts (1992) [5]. They defined the C score as the average number of checkerboard units or motifs between a typical pair of nodes or species. Figs 1A and 2 help explain the concept. Again, the rows of the adjacency matrix A represent species and columns represent islands. In Fig 1A, an example checkerboard motif represents a subset of two species (here, B and D) and two islands (I_3 and I_5). In

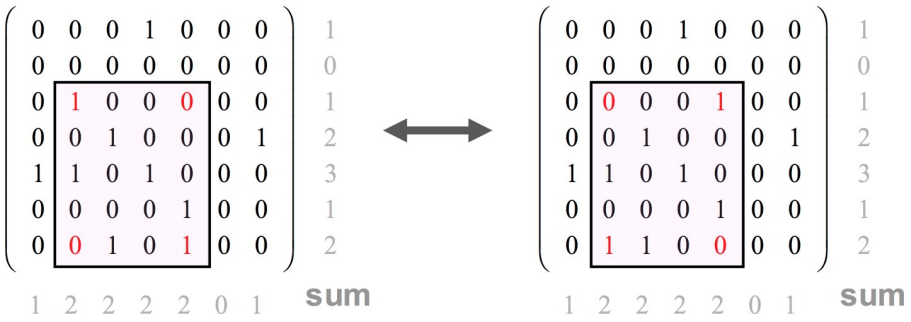

Fig 2. Randomizing matrices with switches. Switches between one checkerboard configuration to another (see 0s and 1s marked in red) leave the row and column sums of the matrix unchanged. One method to generate a set of random samples from the universe of all possible matrices $U(r,c)$ simply requires implementing a large set of switches to randomly chosen checkerboard configurations in the adjacency matrix.

https://doi.org/10.1371/journal.pcbi.1006749.g002

this case, B does not appear on I_3 whereas D does, and although B does appear on I_5, D does not, suggesting competitive interactions. Fig 2 gives an adjacency matrix in which in the first two rows (species-1 and species-2) have no checkerboard motifs, whereas species-3 and species-4 (third and fourth rows) have two checkerboards. The C score of the observed matrix, C_{obs}, is the average number of checkerboard motifs per species pair, when examined for all species pairs. The method then requires finding μ, the average of the C scores in the ensemble of random matrices, and their standard deviation σ. These statistics allow determination of how many standard deviations the observed C score is from the mean, namely,

$$z = (C_{obs} - \mu)/\sigma.$$

A large value of z (e.g., $z > 1.96$) indicates the checkerboard motif is overrepresented in the network, relative to that expected by chance. This is identical to the method for studying overrepresentation of motifs described by Milo and colleagues (2002) [1] more than 10 years later.

Similar ideas were also used even earlier. In the late 1970s, Holland and Leinhardt (1976) [6,27] attempted to identify small-scale social structure using 3-node motifs (triadic structures of Fig 1). Their research, however, was restricted to analyses of specific classes of random matrices (e.g., having row and column sums that are on average all the same constant), rather than samples having the exact network structure $U(r,c)$ particular to r and c. The latter approach allows analysis of a much wider and very flexible range of network topologies. Chase (1980, 1982) [28,29] also used triadic network motifs to study hierarchical relationships in animal societies, but these studies were restricted to specific classes of random matrices rather than samples from $U(r,c)$.

The random matrix and network motif methods, once introduced in ecology, metastasized into a huge literature covering foodweb theory, community null models, and assembly rules. This corpus multiplied at a still greater rate when Milo and colleagues (2002) [1] unleashed this method for use in systems biology. These two different fields share the same technique but were discovered independently. We should not be surprised. The phenomenon of multiple discovery is not a rarity, and their occurrences are not simply strange coincidences. Robert Merton (1961) [19] goes so far as to argue that multiple discoveries, rather than unique ones, are the most common pattern in science, sometimes decades apart. Examples include calculus (Newton and Leibniz), evolution (Darwin and Wallace), and the atomic-bomb (Szilard and Rotblat; see also https://en.wikipedia.org/wiki/List_of_multiple_discoveries).

With respect to binary matrices and the search for motifs indicating nonrandom structure, it is not surprising that independent research should have arisen at the beginning of the 21st

century. Binary matrices and associated graphs have long been subjects of interest in mathematics, tracing back at least as far as Macmahon (1971) [30] and Sukhatme (1938) [31]. A burst of activity by mathematicians in the late 1950s and early 1960s (e.g., [32–35]; cf. [36]) resulted in many theorems about properties of such matrices, including the size of *U(r,c)* and locating a "random" subset of *U(r,c)*. Purely mathematical explorations continued well beyond that period (e.g., [36–39]).

"Null models" arose as a hot topic in community ecology and biogeography in the 1970s, primarily in the context of controversies over the importance of interspecific competition and how such competition would be manifested in geographic distribution patterns [20,40]. Because available data were generally in the form of presence or absence of particular species at particular sites, it was inevitable that ecologists with a mathematical cast of mind would eventually come to represent them as species-by-site binary matrices, and the question of manifestations of interspecific competition would then reduce to seeking submatrices representing pairs of mutually exclusive species—motifs, in network terminology. An early ecological effort by Pielou and Pielou (1968) [41] came close to representing such data as a binary matrix but instead turned to analyzing the data as a contingency table. Connor and Simberloff (1979) [20] instead first used a binary matrix representation and analysis. Because the subject of interspecific competition was prominent and controversial during that period, it was inevitable that methods for examining such matrices proliferated in the literature, originally largely independently of the mathematical literature. At approximately the same time, exploration of social networks became a major research focus in sociology, leading to a similar attempt to define and enumerate *U(r,c)* and to search for nonrandom patterns in observed networks [42,43]. The rise of network theory in several fields then led almost automatically to research on how to identify patterns and thus to the depiction of network motifs (including checkerboard motifs in ecological networks [44]). Because the fields are quite disparate, some of the relevant literature consists of independently inventing the same wheel. As a result, the rediscovery process occurred over several scientific disciplines, both in parallel and out of sync, and over many years as we have outlined here (and as also independently discussed in Fosdick and colleagues (2018) [45] but from the more recent perspective of stub-labeled configuration models). For these reasons, it is not surprising that the general algorithm for detecting network motifs as invented by Milo and colleagues (2002) [1] is almost identical to that developed by Stone (1988) [21] and Stone and Roberts (1992) [5], which in turn has close similarities to the algorithm suggested by Connor and Simberloff (1979) [20] in their study of ecological networks.

References

1. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: simple building blocks of complex networks. Science (80-). 2002; 298: 824–827. https://doi.org/10.1126/science.298.5594.824 PMID: 12399590
2. Milo R, Itzkovitz S, Kashtan N, Levitt R, Shen-Orr S, Ayzenshtat I, et al. Superfamilies of evolved and designed networks. Sci (New York, NY). 2004; 303: 1538–1542. https://doi.org/10.1126/science.1089167 PMID: 15001784
3. Shen-Orr SS, Milo R, Mangan S, Alon U. Network motifs in the transcriptional regulation network of Escherichia coli. Nat Genet. 2002; 31: 64–68. https://doi.org/10.1038/ng881 PMID: 11967538
4. Alon U. Network motifs: Theory and experimental approaches. Nat Rev Genet. 2007; 8: 450–461. https://doi.org/10.1038/nrg2102 PMID: 17510665
5. Stone L, Roberts A. Competitive exclusion, or species aggregation?—An aid in deciding. Oecologia. 1992; 91: 419–424. https://doi.org/10.1007/BF00317632 PMID: 28313551
6. Holland PW, Leinhardt S. Local structure in social networks. Sociol Methodol. 1976; 1–45.
7. Wasserman S, Faust K. Social network analysis: methods and applications. Cambridge, UK, Cambridge University. 1994.

8. House T, Keeling MJ. Epidemic prediction and control in clustered populations. J Theor Biol. Elsevier; 2011; 272: 1–7. https://doi.org/10.1016/j.jtbi.2010.12.009 PMID: 21147131
9. Molina C, Stone L. Modelling the spread of diseases in clustered networks. J Theor Biol. Elsevier; 2012; 315: 110–118. https://doi.org/10.1016/j.jtbi.2012.08.036 PMID: 22982137
10. Lim WA, Lee CM, Tang C. Design Principles of Regulatory Networks: Searching for the Molecular Algorithms of the Cell. Mol Cell. Elsevier Inc.; 2013; 49: 202–212. https://doi.org/10.1016/j.molcel.2012.12.020 PMID: 23352241
11. Chen L, Qu X, Cao M, Zhou Y, Li W, Liang B, et al. Identification of breast cancer patients based on human signaling network motifs. Sci Rep. 2013; 3: 1–7. https://doi.org/10.1038/srep03368 PMID: 24284521
12. Messé A, Hütt MT, Hilgetag CC. Toward a theory of coactivation patterns in excitable neural networks. PLoS Comput Biol. 2018; 14: 1–19. https://doi.org/10.1371/journal.pcbi.1006084 PMID: 29630592
13. Sporns O, Kötter R. Motifs in Brain Networks. PLoS Biol. 2004;2. https://doi.org/10.1371/journal.pbio.0020369 PMID: 15510229
14. Strogatz DJ. WH. Collective Dynamics of "Small-World" Networks. Nature. 1998; 393: 440–442. https://doi.org/10.1038/30918 PMID: 9623998
15. Saracco F, Di Clemente R, Gabrielli A, Squartini T. Detecting early signs of the 2007–2008 crisis in the world trade. Sci Rep. Nature Publishing Group; 2016; 6: 1–11. https://doi.org/10.1038/srep30286 PMID: 27461469
16. Saracco F, Di Clemente R, Gabrielli A, Squartini T. Randomizing bipartite networks: The case of the World Trade Web. Sci Rep. Nature Publishing Group; 2015; 5: 1–18. https://doi.org/10.1038/srep10595 PMID: 26029820
17. Jiang S, Fiore GA, Yang Y, Ferreira JJ, Frazzoli E, González MC. A Review of Urban Computing for Mobile Phone Traces: Current Methods, Challenges and Opportunities. Proc 2nd ACM SIGKDD Int Work Urban Comput. 2013;
18. Ansariola M, Megraw M, Koslicki D. IndeCut evaluates performance of network motif discovery algorithms. Bioinformatics. 2018; 34: 1514–1521. https://doi.org/10.1093/bioinformatics/btx798 PMID: 29236975
19. Merton RK. ingletons and multiples in scientific discovery: A chapter in the sociology of science. Proc Am Philos Soc. 1961; 13: 470–86.
20. Connor EF, Simberloff D. The Assembly of Species Communities: Chance or Competition? Ecology. 1979; 60: 1132–1140.
21. Stone L. Some problems of community ecology processes, patterns and species persistence in ecosystems. 1988. Available from: https://figshare.com/articles/Some_problems_of_community_ecology_processes_patterns_and_species_persistence_in_ecosystems/4519409
22. Simberloff D. Analysis of presence/absence data for species on islands: Passerine birds of the Cyclades. Biol Gall. 1986; 43–68.
23. Zaman A, Simberloff D. Random binary matrices in biogeographical ecology—instituting a good neighbor policy. Environ Ecol Stat. 2002; 9: 405–421. https://doi.org/10.1023/A:1020918807808
24. Artzy-Randrup Y, Stone L. Generating uniformly distributed random networks. Phys Rev E—Stat Nonlinear Soft Matter Phys. American Physical Society; 2005; 72: 056708. https://doi.org/10.1103/PhysRevE.72.056708 PMID: 16383786
25. Miklós I, Podani J. Randomization of presence-absence matrics: comments and new algorithms. Ecology. 2004; 85: 86–92. https://doi.org/10.1890/03-0101
26. Strona G, Nappo D, Boccacci F, Fattorini S, San-Miguel-Ayanz J, Gotelli NJ, et al. A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals. Nat Commun. Nature Publishing Group; 2014; 5: 2606–2621. https://doi.org/10.1038/ncomms5114 PMID: 24916345
27. Holland PW, Leinhardt S. A Method for Detecting Structure in Sociometric Data. Soc Networks. Academic Press; 1977; 411–432. https://doi.org/10.1016/B978-0-12-442450-0.50028–6
28. Chase ID. Social Process and Hierarchy Formation in Small Groups: A Comparative Perspective. Am Sociol Rev. 1980; 45: 905. https://doi.org/10.2307/2094909
29. Chase ID. Dynamics of hierarchy formation: the sequential development of dominance relationships. Behaviour. 1982; 80: 218–239. https://doi.org/10.1163/156853982X00364
30. Macmahon PA. Combinatory Analysis. Cambridge University Press; 1971.
31. Sukhatme PV. On bipartitional functions. Philos Trans R Soc London Ser A, Math Phys Sci. 1938; 237: 375–409.
32. Gale D. A theorem on flows in networks. Pacific J Math. 1957; 7.

33. Ryser HJ. Combinatorial properties of matrices of zeros and ones. Canad J Math. 1957; 9: 371–377.
34. Haber RM. Term rank of 0,l matrices. Rend Sem Mat Univ Padova. 1960; 30: 24–51.
35. Fulkerson DR, Ryser HJ. Multiplicities and minimal widths for (0,l)-matrices. Canad J Math. 1962; 14: 498–508.
36. Brualdi RA. Matrices of zeros and ones with fixed row and column sum vectors. Linear Algebra Appl. 1980; 33: 159–231.
37. Snapper E. Group characters and non-negative integral matrices. J Algebr. 1971; 19: 520–535.
38. Bender EA. The asymptotic number of non-negative integer matrices with given row and column sums. Discrete Math. 1974; 10: 217–223.
39. Wang BY. Precise number of (0,1)-matrices in A(R,S). Sci Sin Ser A. 1988; 31: 1–6.
40. Gotelli NJ, Graves GR. Null models in ecology. Washington, D.C., USA: Smithsonian Institution Press; 1996.
41. Pielou DP, Pielou EC. Association among species of infrequent occurrence: the insect and spider fauna of *Polyporus betulinus* (Bulliard) Fries. J Theor Biol. 1968; 21: 202–216. PMID: 5700435
42. Rao A. R., Bandyopadhyay S. Measures of reciprocity in a social network. Sankhya, Ser A. 1987; 49: 141–188.
43. Snijders TAB. Enumeration and simulation methods for 0–1 matrices with given marginals. Psychometrika. 1991; 56: 397–417.
44. Toju H, Yamamoto S, Tanabe AS, Hayakawa T, Ishii HS. Network modules and hubs in plant-root fungal biomes. https://doi.org/10.1098/rsif.2015.1097 PMID: 26962029
45. Fosdick BK, Larremore DB, Nishimura J, Ugander J. Configuring Random Graph Models with Fixed Degree Sequences. SIAM Rev. 2018; 60: 315–355. https://doi.org/10.1137/16M1087175