

GraphCrunch: A Tool for Large Network Analyses

Tijana Milenković¹, Jason Lai¹, Nataša Pržulj^{1,*}

1. Department of Computer Science, University of California, Irvine, CA 92697-3435, USA

* Corresponding author (E-mail: natasha@ics.uci.edu)

Abstract

Modeling biological networks is a vibrant research area. A good model should generate graphs that resemble real data as closely as possible across a wide range of statistical measures. Testing the efficacy of a model entails comparing model-derived random graphs to real data. GraphCrunch is an *open-source* software tool that implements the *latest* research on biological network models and properties: it compares real-world networks against a series of random graph models with respect to a variety of network properties.

GraphCrunch currently supports *five* different types of random graphs: Erdős-Rényi random graphs (“ER”) [1], generalized random graphs with the same degree distribution as the real-world network (“ER-DD”) [2], Barabási-Albert type scale-free networks (“SF”) [3], n -dimensional geometric random graphs, where n is any positive integer (“GEO- n D”) [4], and stickiness-index-based model networks (“STICKY”) [5]. Comparing large networks relies on heuristics, such as global and local network properties, due to NP-completeness of the underlying subgraph isomorphism problem. GraphCrunch currently supports *two* computationally expensive “graphlet”-based similarity measures of *local* network structure, where *graphlets* are small connected non-isomorphic induced subgraphs of large networks [6] (see Figure 1). These two local network properties are: relative graphlet frequency distance (“RGF-distance”) [6] and graphlet degree distribution agreement (“GDD-agreement”) [7]. GraphCrunch also computes *five* standard *global* network properties: degree distributions, distributions of shortest path lengths, average network diameters, average clustering coefficients, and clustering spectra.

GraphCrunch *automates* the process of generating random networks drawn from user specified random graph models and evaluating the fit of the network models to a real-world network with respect to global and local network properties. In a single command, GraphCrunch performs all of the following tasks: 1) computes user specified global and local properties of an input real-world network, 2) creates a user specified number of random networks belonging to user specified random graph models, 3) compares how closely each model network reproduces a range of global and local properties (specified in point 1 above) of the real-world network, and 4) produces the statistics of network property similarities between the data and the model networks. GraphCrunch is capable of producing these statistics both in the tabular and in graphical format. An example of visualized output that illustrates the fit of five network models to three data sets with respect to GDD-agreement is presented in Figure 2.

GraphCrunch is the *first* software tool that *compares* real-world networks against a *series* of network models with respect to a *variety* of local and global network properties. It is *easily extendible* to include additional network models and properties. GraphCrunch has *parallel computing capabilities* – it allows for a user specified list of machines on which to distribute computationally expensive searches for local network properties. This is the first network analysis software tool that has built-in parallel computing capabilities, a feature that will become crucial as biological network data sets grow. We introduce GraphCrunch as a comprehensive, parallelizable, and easily extendible open-source research software tool for analyzing and modeling large real-world networks.

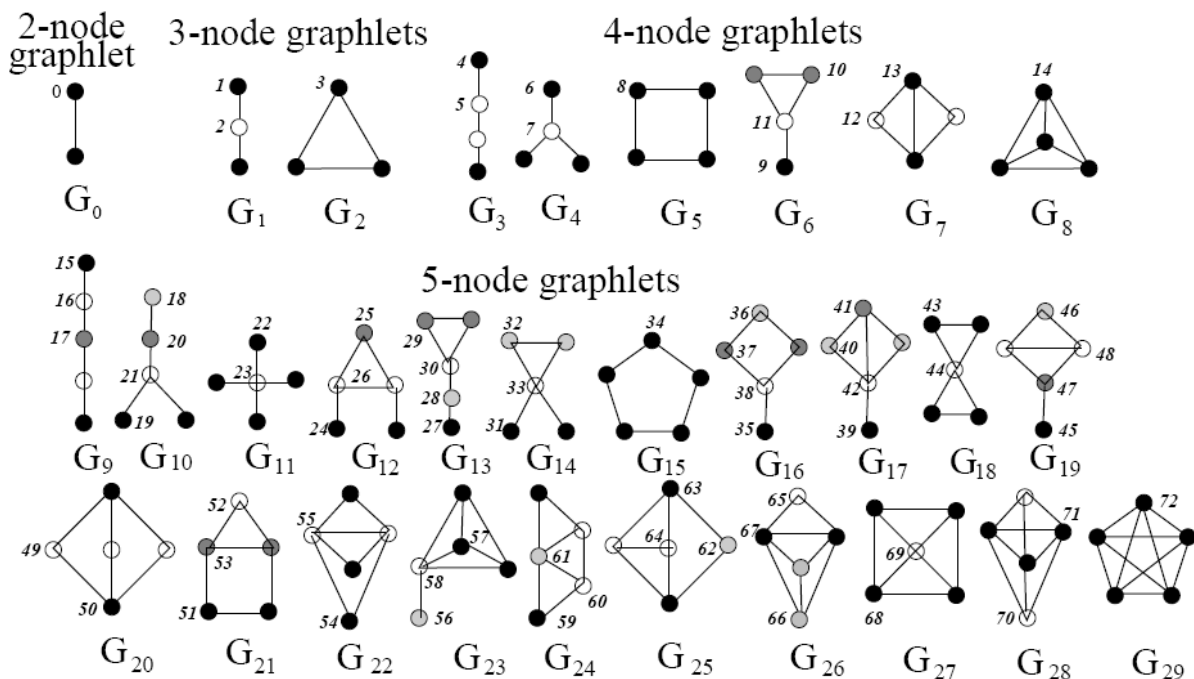


Figure 1: Automorphism orbits 0, 1, 2, ..., 72 for the thirty 2-, 3-, 4-, and 5-node graphlets G_0, G_1, \dots, G_{29} . In a graphlet $G_i, i \in \{0, 1, \dots, 29\}$, nodes belonging to the same orbit are of the same shade. These are used for computing RGF-distance and GDD-agreement. See [6] and [7] for details.

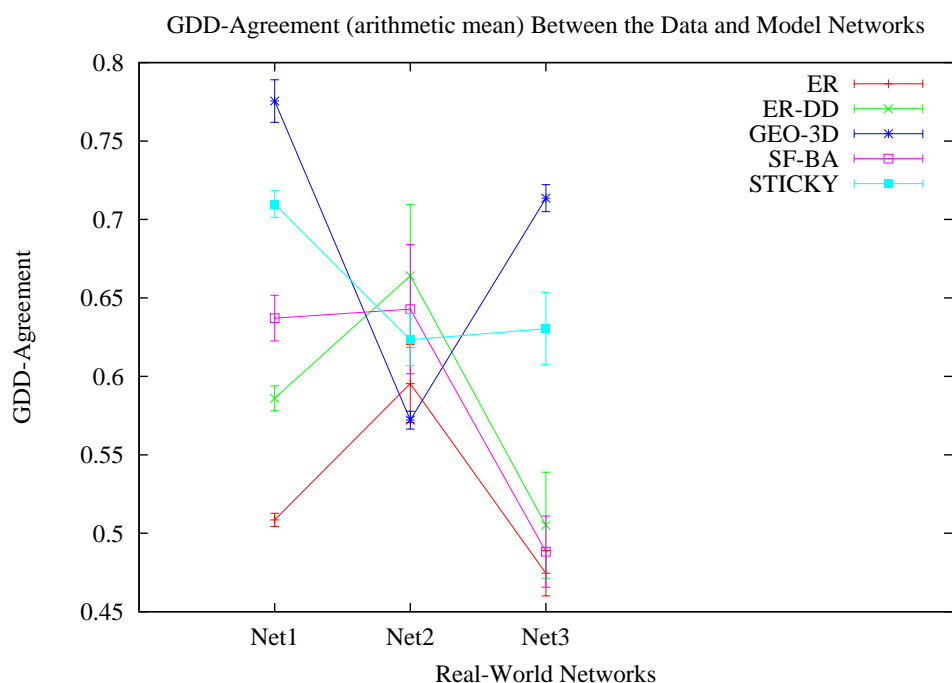


Figure 2: An example of plot created by GraphCrunch that illustrates the fit of five network models (ER, ER-DD, GEO-3D, SF-BA, and STICKY) to three data sets (Net1, Net2 and Net3) with respect to GDD-agreement [7]. Points in the panel represent averages of GDD-agreements over model networks belonging to the same random graph model; the error bars represent one standard deviation above and below the average.

References

- [1] Erdős, P. and Rényi, A. (1959) On random graphs. *Publicationes Mathematicae*, **6**, 290–297.
- [2] Molloy, M. and Reed, B. (1995) A critical point of random graphs with a given degree sequence. *Random Structures and Algorithms*, **6**, 161–180.
- [3] Barabási, A.-L. and Albert, R. (1999) Emergence of scaling in random networks. *Science*, **286**(5439), 509–512.
- [4] Penrose, M. (2003) *Geometric Random Graphs*, Oxford University Press.
- [5] Pržulj, N. and Higham, D. (2006) Modelling proteinprotein interaction networks via a stickiness index. *Journal of the Royal Society Interface*, **3**(10), 711–716.
- [6] Pržulj, N., Corneil, D. G., and Jurisica, I. (2004) Modeling interactome: Scale-free or geometric?. *Bioinformatics*, **20**(18), 3508–3515.
- [7] Pržulj, N. (2006) Biological network comparison using graphlet degree distribution. *Bioinformatics*, **23**, e177–e183.