Journal of Statistical Physics, Vol. 128, Nos. 1/2, July 2007 (© 2007)
DOI: 10.1007/s10955-006-9184-x

The Critical Point of *k*-Clique Percolation in the Erdős–Rényi Graph

Gergely Palla,[1,2] Imre Derényi[2] and Tamás Vicsek[1,2]

Received February 1, 2006; accepted July 18, 2006
Published Online August 1, 2006

Motivated by the success of a *k*-clique percolation method for the identification of overlapping communities in large real networks, here we study the *k*-clique percolation problem in the Erdős–Rényi graph. When the probability p of two nodes being connected is above a certain threshold $p_c(k)$, the complete subgraphs of size k (the k-cliques) are organized into a giant cluster. By making some assumptions that are expected to be valid below the threshold, we determine the average size of the k-clique percolation clusters, using a generating function formalism. From the divergence of this average size we then derive an analytic expression for the critical linking probability $p_c(k)$.

KEY WORDS: random graph, network, percolation, community.

1. INTRODUCTION

Many complex systems in nature and society can be successfully represented in terms of networks capturing the intricate web of connections among the units they are made of. Graphs corresponding to these real networks exhibit unexpected non-trivial properties, e.g., new kinds of degree distributions, anomalous diameter, spreading phenomena, clustering coefficient, and correlations.[1–5] In recent years, there has been a quickly growing interest in the local structural units of networks. Small and well defined subgraphs consisting of a few vertices have been introduced as "motifs".[6] Their distribution and clustering properties[6–8] can be considered as important global characteristics of real networks. Somewhat larger units, associated with more highly interconnected parts[9–21] are usually called

[1] Biological Physics Research Group of HAS, Pázmány P. stny. 1A, H-1117 Budapest, Hungary.
[2] Department of Biological Physics, Eötvös University, Pázmány P. stny. 1A, H-1117 Budapest, Hungary.

clusters, communities, cohesive groups, or modules, with no widely accepted, unique definition. Such building blocks (functionally related proteins,[22,23] industrial sectors,[24] groups of people,[17,25] cooperative players,[26,27] etc.) can play a crucial role in the structural and functional properties of the networks involved. The presence of communities is also a relevant and informative signature of the hierarchical nature of complex systems.[22,28,29]

Most of the methods used for the identification of communities rely on dividing the network into smaller pieces. The biggest drawback of these methods is that they do not allow *overlapping* for the communities. On the other hand, the communities in a complex system are often not isolated from each other, but rather, they have overlaps, e.g., a protein can be part of more than one functional unit,[30] and people can be members in different social groups at the same time.[31] One possibility to overcome this problem is to use a community definition based on k-clique percolation.[32,33] In this approach the communities are associated with k-clique percolation clusters, and can overlap with each other. The communities of large real networks obtained with this method were shown to have significant overlaps, and the statistical properties of the communities exhibited non-trivial universal features.[33]

In this manuscript we focus on the basic properties of k-clique percolation. In a recent work we have already proposed an expression for the critical point of the k-clique percolation in the Erdős–Rényi (E–R) graph using simple heuristic arguments.[32] This expression has also been supported by our numerical simulations. The goal of this manuscript is to make these result stronger by providing a more detailed analytical derivation using only a few reasonable assumptions, expected to be valid below the critical point. We note that the critical point of k-clique percolation plays a crucial role in the community finding as well. When dealing with a network containing weighted links, one can introduce a weight threshold and exclude links weaker than the threshold from the investigation. When the threshold is very high, only a few disintegrated community remains, whereas in case of a very low threshold, a giant community arises smearing out the details of the community structure by merging (and making invisible) many smaller communities. To find a community structure as highly structured as possible, one needs to set the threshold close to the critical point of the k-clique percolation.

2. K-CLIQUE PERCOLATION IN THE E–R GRAPH

In the field of complex networks, the classical E–R uncorrelated random graph[34] serves both as a test bed for checking all sorts of new ideas concerning networks in general, and as a prototype of random graphs to which all other random graphs can be compared. One of the most conspicuous early result on the E–R graphs was related to the percolation transition taking place at $p = p_c \equiv 1/N$, where p is the probability that two vertices are connected by an edge and N is

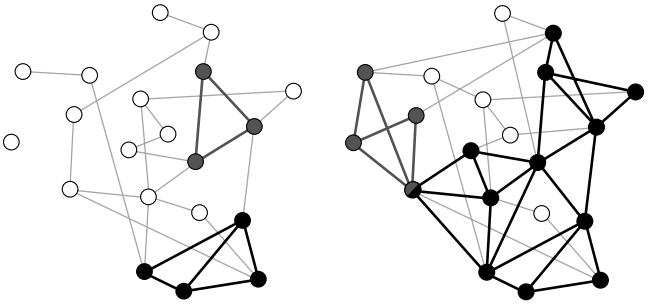

Fig. 1. Sketches of two E–R graphs of $N = 20$ vertices and with edge probabilities $p = 0.13$ (left one) and $p = 0.22$ (right one, generated by adding more random edges to the left one). In both cases all the edges belong to a "giant" connected component, because the edge probabilities are much larger than the threshold ($p_c \equiv 1/N = 0.05$) for the classical E–R percolation transition. However, in the left one p is below the 3-cliques percolation threshold, $p_c(3) \approx 0.16$, calculated from Eq. (19), therefore, only a few scattered 3-cliques (triangles) and small 3-clique percolation clusters (distinguished by black and dark gray edges) can be observed. In the right one, on the other hand, p is above this threshold and, as a consequence, most 3-cliques accumulate in a "giant" 3-clique percolation cluster (black edges). This graph also illustrates the overlap (half black, half dark gray vertex) between two clusters (black and dark gray).

the total number of vertices in the graph. The appearance of a *giant component* in a network, which is also referred to as the *percolating component*, results in a dramatic change in the overall topological features of the graph and has been in the center of interest for other networks as well.

In this manuscript we address the general question of subgraph percolation in the E–R model. In particular, we provide an analytic expression for the threshold probability at which the percolation transition of complete subgraphs of size k (the k-cliques) takes place. Before proceeding we need to go through some basic definitions:

- *k-clique*: a complete (fully connected) subgraph of k vertices.[35]
- *k-clique adjacency*: two k-cliques are adjacent if they share $k - 1$ vertices, *i.e.*, if they differ only in a single node.
- *k-clique chain*: a subgraph, which is the union of a sequence of adjacent k-cliques.
- *k-clique connectedness*: two k-cliques are k-clique-connected if they are parts of a k-clique chain.
- *k-clique percolation cluster (or component)*: a maximal k-clique-connected subgraph, *i.e.*, it is the union of all k-cliques that are k-clique-connected to a particular k-clique.

The above concept of k-clique percolation is illustrated in Fig. 1, where both graphs contain two 3-clique percolation clusters, the smaller ones in dark gray

and the larger ones in black. We note that these objects can be considered as interesting specific cases of the general graph theoretic objects defined by Everett and Borgatti[36] and by Batagelj and Zaversnik[37] in very different contexts.

An illustration of the k-clique percolation clusters can be given by "*k-clique template rolling*". A k-clique template can be thought of as an object that is isomorphic to a complete graph of k nodes. Such a template can be placed onto any k-clique of the network, and rolled to an adjacent k-clique by relocating one of its nodes and keeping its other $k-1$ nodes fixed. Thus, the k-clique-communities of a graph are all those subgraphs that can be fully explored by rolling a k-clique template in them but cannot be left by this template. We note that a k-clique percolation cluster is very much like a regular edge percolation cluster in the *k-clique adjacency graph*, where the vertices represent the k-cliques of the original graph, and there is an edge between two vertices if the corresponding k-cliques are adjacent. Moving a particle from one vertex of this adjacency graph to another one along an edge is equivalent to rolling a k-clique template from one k-clique of the original graph to an adjacent one.

3. THE GENERATING FUNCTIONS

In our investigation of the critical point of the k-clique percolation in the E–R graph we shall rely on the generating function formalism in a fashion similar to that of Ref. 38. Therefore, in this section we first summarize the definition and the most important properties of the generating functions. If a random variable ξ can take non-negative integer values according to some probability distribution $\mathcal{P}(\xi = n) \equiv \rho(n)$, then the corresponding generating function is given by

$$G_\rho(x) \equiv \langle x^\xi \rangle = \sum_{n=0}^{\infty} \rho(n) x^n. \tag{1}$$

The generating-function of a properly normalized distribution is absolute convergent for all $|x| \leq 1$ and hence has no singularities in this region. For $x = 1$ it is simply

$$G_\rho(1) = \sum_{n=0}^{\infty} \rho(n) = 1. \tag{2}$$

The original probability distribution and its moments can be obtained from the generating-function as

$$\rho(n) = \frac{1}{n!} \left. \frac{d^n G_\rho(x)}{dx^n} \right|_{x=0}, \tag{3}$$

$$\langle \xi^l \rangle = \sum_{n=0}^{\infty} n^l \rho(n) = \left[\left(x \frac{d}{dx} \right)^l G_\rho(x) \right]_{x=1}. \tag{4}$$

And finally, if $\eta = \xi_1 + \xi_2 + \cdots + \xi_l$, where $\xi_1, \xi_2, \ldots, \xi_l$ are independent random variables (with non-negative integer values), then the generating function corresponding to $\mathcal{P}(\eta = n) \equiv \sigma(n)$ is given by

$$\begin{aligned} G_\sigma(x) &= \langle x^\eta \rangle = \langle x^{\xi_1} x^{\xi_2} \cdots x^{\xi_l} \rangle = \langle x^{\xi_1} \rangle \langle x^{\xi_2} \rangle \cdots \langle x^{\xi_l} \rangle \\ &= G_{\rho_1}(x) G_{\rho_2}(x) \cdots G_{\rho_l}(x). \end{aligned} \tag{5}$$

4. THE CRITICAL POINT

In this section we arrive at the derivation of the critical point of the k-clique percolation in the E–R graph in the $N \to \infty$ limit. We begin by considering the probability distribution $r(n)$ of the number of k-cliques adjacent to a randomly selected k-clique. Finding a k-clique B adjacent to a selected k-clique A is equivalent to finding a node outside A linked to at least $k-1$ nodes in A. The number of possibilities for this node is $N-k$. Links in the E–R graph are independent of each other, therefore the probability that a given node is linked to all nodes in A is p^k, whereas the probability that it is linked to $k-1$ nodes in A is $k(1-p)p^{k-1}$. Therefore, to leading order in N the average number of k-cliques adjacent to a randomly selected one is

$$\langle r \rangle = (N-k)[k(1-p)p^{k-1} + p^k] \simeq Nkp^{k-1}. \tag{6}$$

From the independence of the links it also follows that the probability distribution $r(n)$ becomes Poissonean, which can be written as

$$r(n) = \exp(-Nkp^{k-1}) \frac{(Nkp^{k-1})^n}{n!}. \tag{7}$$

Let us suppose that we are below the percolation threshold, therefore, k-cliques are rare, adjacent k-cliques are even more rare, and loops in the k-clique adjacency graph are so rare that we can assume it to be tree-like.[39] In this case the size of a connected component in the k-clique adjacency graph (corresponding to a k-clique percolation cluster) can be evaluated by counting the number of k-cliques reached in an "invasion" process as follows. We start at an arbitrary k-clique in the component, and in the first step we invade all its neighbors in the k-clique adjacency graph. From then on, whenever a k-clique is reached, we proceed by invading all its neighbors, except for the one the k-clique has been reached from, as shown in Fig. 2a. In terms of the original graph, this is equivalent to rolling a k-clique template to all adjacent k-cliques except for the one we arrived from in the previous step.

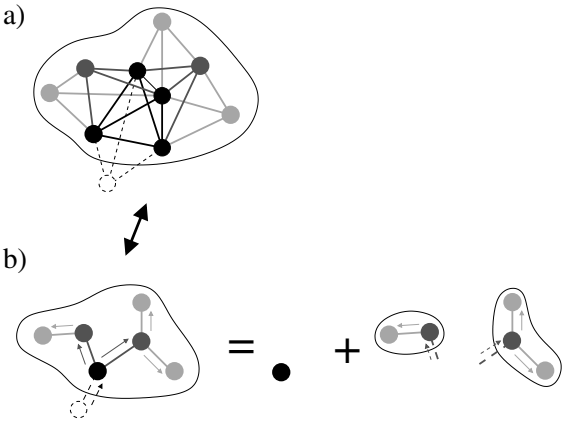

Fig. 2. Schematic picture of the evaluation of the size of a k-clique percolation cluster by counting the number of k-cliques reached in an "invasion" process. (a) Let us suppose that we arrived at the black colored k-clique from the k-clique marked by dashed lines. In the next step we must proceed to the k-cliques shown in dark gray, and then finally to the k-cliques marked in light gray. (b) The corresponding k-clique adjacency graph is shown on the right. The size of the connected component in the k-clique adjacency graph we can invade from the black k-clique (by excluding the link through which we initially reached it) is equal to one plus the sum of the sizes of the connected components invaded from the dark gray k-cliques in the same way.

In the invasion process described above, we can assign to each k-clique the subgraph in the k-clique percolation cluster that was invaded from it. (Note that we assumed the k-clique adjacency graph to be tree-like). Let us denote by $I(n)$ the probability, that the subgraph reached from an arbitrary starting k-clique in the invasion process contains n number of k-cliques, including the starting k-clique as well. This subgraph is actually equal to a k-clique percolation cluster. Similarly, let $H(n)$ denote the probability that the subgraph invaded from a k-clique appearing later in the invasion process (i.e., from a k-clique that is not the starting one) contains n number of k-cliques. This is equivalent to the probability that by starting at a randomly selected k-clique and trying to roll a k-clique template via all possible subsets of size $k-1$ except for one, then by succeedingly rolling the template on and on, in all possible directions without turning back, a k-clique percolation "branch" of size n is invaded. And finally, let $H_m(n)$ be the probability, that if pick m number of k-cliques randomly, then the sum of the sizes of the k-clique branches that we can invade in this way consists of n number of k-cliques. Since we are below the percolation threshold, the k-clique adjacency graph consists of many dispersed components of small size, and the probability that two (or more) k-cliques out of m belong to the same k-clique percolation cluster is negligible. Hence, according to Eq. (5), the generating functions corresponding to $H(n)$ and

$H_m(n)$, denoted by $G_H(x)$ and $G_{H_m}(x)$ respectively are related to each other as:

$$G_{H_m}(x) = [G_H(x)]^m . \tag{8}$$

Let $q(n)$ denote the probability, that for a randomly selected k-clique, by excluding one of its possible subsets of size $k-1$, we can roll a k-clique template through the remaining subsets to n adjacent k-cliques. This distribution is very similar to $r(n)$, except that in this case we can use only $k-1$ subsets instead of k in the k-clique to roll the k-clique template further, therefore

$$q(n) = \exp(-N(k-1)p^{k-1})\frac{(N(k-1)p^{k-1})^n}{n!}. \tag{9}$$

By neglecting the loops in the k-clique adjacency graph, H_n can be expressed as

$$H(n) = q(0)H_0(n-1) + q(1)H_1(n-1) + q(2)H_2(n-1) + \ldots, \tag{10}$$

as explained in Fig. 2b. By taking the generating function of both sides and using Eqs. (3) and (8), we obtain

$$\begin{aligned} G_H(x) &= \sum_{n=0}^{\infty}\left[\sum_{m=0}^{\infty} q(m)H_m(n-1)\right]x^n \\ &= \sum_{n=0}^{\infty}\left[\sum_{m=0}^{\infty} q(m)\frac{1}{(n-1)!}\,\frac{d^{n-1}}{dx^{n-1}}\,[G_H(x)]^m\Big|_{x=0}\right]x^n \\ &= \sum_{m=0}^{\infty} q(m)\,[G_H(x)]^m\,x = xG_q(G_H(x)), \end{aligned} \tag{11}$$

where $G_q(x)$ denotes the generating function of the distribution $q(n)$.

We can write an equation similar to Eq. (10) for $I(n)$ as well, in the form of

$$I(n) = r(0)H_0(n-1) + r(1)H_1(n-1) + r(2)H_2(n-1) + \cdots \tag{12}$$

Again, by taking the generating functions of both sides we arrive at

$$G_I(x) = xG_r(G_H(x)), \tag{13}$$

where $G_r(x)$ denotes the generating function of $r(n)$. From Eqs. (4) and (13) we get

$$\langle I\rangle = G_I'(1) = G_r(G_H(1)) + G_r'(G_H(1))G_H'(1) = 1 + G_r'(1)G_H'(1) \tag{14}$$

for the average size of the components invaded from a randomly selected k-clique. Using Eq. (11) we can write

$$G_H'(1) = G_q(G_H(1)) + G_q'(G_H(1))G_H'(1) = 1 + G_q'(1)G_H'(1), \tag{15}$$

from which $G'_H(1)$ can be expressed as

$$G'_H(1) = \frac{1}{1 - G'_q(1)}. \tag{16}$$

By substituting this back into Eq. (14) we get

$$\langle I \rangle = 1 + \frac{G'_r(1)}{1 - G'_q(1)} = 1 + \frac{\langle r \rangle}{1 - \langle q \rangle}. \tag{17}$$

The above expression for the expected size of the connected components in the k-clique adjacency graph invaded from a randomly selected k-clique diverges when

$$\langle q \rangle = N(k-1)p^{k-1} = 1. \tag{18}$$

This point marks the phase transition at which a giant component (corresponding to a giant k-clique percolation cluster) first appears. Therefore, our final result for the critical linking probability for the appearance of the giant component is

$$p_c(k) = \frac{1}{[N(k-1)]^{\frac{1}{k-1}}}. \tag{19}$$

This result reassures the findings of[32] based on heuristic arguments and numerical simulations. Obviously, for $k = 2$ our result agrees with the known percolation threshold ($p_c = 1/N$) for E–R graphs, because 2-clique connectedness is equivalent to regular (edge) connectedness.

5. CONCLUSIONS

The phenomenon of k-clique percolation provides an effective tool for finding overlapping communities in large networks. In this article we derived the critical linking probability for the E–R graph in the $N \to \infty$ limit. Our method involved the use of generating functions and was based on the assumption that up to the critical point, loops in the k-clique adjacency graph are negligible. Our findings are in complete agreement with earlier results based on heuristic arguments and numerical simulations.

ACKNOWLEDGMENTS

This work has been supported in part by the Hungarian Science Foundation (OTKA), grant Nos. F047203 and T049674.

REFERENCES

1. D. J. Watts and S. H. Strogatz, *Nature* **393**: 440 (1998).
2. A.-L. Barabási and R. Albert, *Science* **286**: 509 (1999).
3. R. Albert and A.-L. Barabási, *Rev. Mod. Phys.* **74**: 47 (2002).
4. J. F. F. Mendes and S. N. Dorogovtsev, *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford University Press, Oxford, 2003).
5. A. Barrat, M. Barthelemy and A. Vespignani, *Phys. Rev. Lett.* **92**: 228701 (2004).
6. R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii and U. Alon, *Science* **298**: 824 (2002).
7. A. Vazquez, R. Dobrin, D. Sergi, J.-P. Eckmann, Z. Oltvai and A.-L. Barabási, *Proc. Natl. Acad. Sci. USA* **101**: 17945 (2004).
8. J.-P. Onnela, J. Saramäki, J. Kertész and K. Kaski, *Phys. Rev. E* **71**: 065103 (2005).
9. M. Blatt, S. Wiseman and E. Domany, *Phys. Rev. Lett.* **76**: 3251 (1996).
10. M. Girvan and M. E. J. Newman, *Proc. Natl. Acad. Sci. USA* **99**: 7821 (2002).
11. H. Zhou, *Phys. Rev. E* **67**: 061901 (2003).
12. M. E. J. Newman, *Phys. Rev. E* **69**: 066133 (2004).
13. F. Radicchi, C. Castellano, F. Cecconi, V. Loreto and D. Parisi, *Proc. Natl. Acad. Sci. USA* **101**: 2658 (2004).
14. L. Donetti and M. A. Muñoz, *J. Stat. Mech.* (2004) P10012.
15. D. M. Wilkinson and B. A. Huberman, *Proc. Natl. Acad. Sci. USA* **101**: 5241 (2004).
16. J. Reichardt and S. Bornholdt, *Phys. Rev. Lett.* **93**: 218701 (2004).
17. J. Scott, *Social Network Analysis: A Handbook*, 2nd ed. (Sage Publications, London, 2000).
18. R. M. Shiffrin and K. Börner, Mapping knowledge domains. *Proc. Natl. Acad. Sci. USA* **101**(Suppl. 1): 5183 (2004).
19. B. S. Everitt, *Cluster Analysis*, 3th ed. (Edward Arnold, London, 1993).
20. S. Knudsen, *A Guide to Analysis of DNA Microarray Data*, 2nd ed. (Wiley-Liss, 2004).
21. M. E. J. Newman, Detecting community structure in networks. *Eur. Phys. J. B.* **38**: 321 (2004).
22. E. Ravasz, A. L. Somera, D. A. Mongru, Z. Oltvai and A.-L. Barabási, *Science* **297**: 1551 (2002).
23. V. Spirin and L. A. Mirny, *Proc. Natl. Acad. Sci. USA* **100**: 12123 (2003).
24. J. P. Onnela, A. Chakraborti, K. Kaski, J. Kertész and A. Kanto, *Phys. Rev. E* **68**: 056110 (2003).
25. D. J. Watts, P. S. Dodds and M. E. J. Newman, *Science* **296**: 1302 (2002).
26. J. Vukov and Gy. Szabó, *Phys. Rev. E* **71**: 036133 (2005).
27. Gy. Szabó, J. Vukov and A. Szolnoki, *Phys. Rev. E* **72**: 047107 (2005).
28. T. Vicsek, *Nature* **418**: 131 (2002).
29. R. Guimerá, L. Danon, A. Díaz-Guilera, F. Giralt and A. Arenas, *Phys. Rev. E* **68**: 065103 (2003).
30. A. C. Gavin, *et al.*, *Nature* **415**: 141 (2002).
31. P. Eds Carrington, J. Scott and S. Wasserman, *Models and Methods in Social Network Analysis*, Ch. 7, K. Faust ed. (Cambridge University Press, New York, 2005).
32. I. Derényi, G. Palla and T. Vicsek, *Phys. Rev. Lett.* **94**: 160202 (2005).
33. G. Palla, I. Derényi, I. Farkas and T. Vicsek, *Nature* **435**: 814 (2005).
34. P. Erdős and A. Rényi, *Publ. Math. Inst. Hung. Acad. Sci.* **5**: 17 (1960).
35. B. Bollobás, *Random graphs*, 2nd ed. (Cambridge University Press, Cambridge, 2001).
36. M. G. Everett and S. P. Borgatti, *Connections* **21**: 49 (1998).
37. V. Batagelj and M. Zaversnik, arXiv: cs.DS/0308011 (2003).
38. M. E. J. Newman, S. H. Strogatz and D. J. Watts, *Phys. Rev. E* **64**: 026118 (2001).
39. This assumption is an approximation since the adjacency graph is weakly assortative.