SAND2014-1830C

FINDING A PLANTED CLIQUE IN A DISTRIBUTED SOCIAL NETWORK

Jonathan W. Berry, Aaron Kearns, Cynthia Phillips, Jared Saia

SIAM Workshop on Network Science 2014
July 6-7 · Chicago

Summary

We consider the problem of finding a planted clique in a social network whose edges are distributed between two players. We make two empirically-justified assumptions about the social network: 1) the maximum degree is $\sqrt{n}$ and 2) the clustering coefficient of a node with degree d is $O(1/d^2)$, a property implied with high probability from the log-normal distribution posited by Kolda et. al. [7]. Our goal is to find the planted clique while minimizing the amount of communication between the players.

We give a protocol that provably ensures whp that both players find the clique, while requiring at most polylogarithmic communication, and polynomial computation. We believe that our algorithm can be generalized to the case where edges are distributed among a constant number of players and/or to finding a planted γ-quasi-clique.

Introduction

Bob and Alice are two separate entities who observe the world, learning about relationships, each building a graph. The graphs are potentially huge, gathered over years, and may represent considerable financial investment. Bob and Alice observe the same world, but may observe with different emphasis or different methods. Although each knows nothing about the other's graph, we assume they use the same naming conventions for the nodes. Thus, if Alice and Bob both have a node for object x, they use the same name, and if Bob sends Alice an edge with x as an endpoint, she knows it is adjacent to her node labeled x.

Bob and Alice would like to cooperate to answer questions about the union of their graphs. However, we assume communication between Alice and Bob is constrained to be polylogarithmic. This may be because of the economic value of information, internal corporate policy, or simply limited bandwidth.

We make two key assumptions about the social network that Alice and Bob observe. The first is that the maximum degree of any node is $\sqrt{n}$. This can be enforced with a log-normal distribution [7], and has been assumed in other social-graph analysis such as Britton et. al. [1] and Chung-Lu [3].

Our second assumption concerns how the clustering coefficient of a node can depend on the node degree. In particular, we assume the clustering coefficient of a node with degree d is $O(1/d^2)$. Empirical evidence suggests this assumption is reasonable for many social networks [7]. This is especially true when social networks are "cleaned" of abnormal behavior such as twitter accounts that reciprocate all follower relationships [9].

Moreover, two sociological properties may explain this phenomena. First, the *Strong Triadic Closure* property of [6] says that if node x has strong links with nodes y and z then there is likely to be a (potentially weak) link between y and z. Second, sociologists have argued that the number of strong links that a node can have in a social network is bounded [5]. These two facts together suggest that the clustering coefficient must drop off at least as fast as an inverse quadratic in the degree of the node.

Let n be the number of nodes in this social network. For our theoretical results, we assume that Alice and Bob both know n. Our experimental results use a slightly revised algorithm that does not require this knowledge.

We assume that a subset of $O(\ln n)$ nodes are chosen uniformly at random and that edges are added among these nodes to form a clique. We can extend these results for an adversarial placement of the clique[1]. We define $G = (V, E)$ to be the base social network with the planted clique. The adversary distributes the edges arbitrarily to Alice and Bob subject to the constraint that at least one player knows each planted clique edge. Some edges of the base graph may be in neither graph, but this only makes the problem easier.

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

[1] The extended analysis requires the generalized log-normal degree distribution. Kolda et. al. [7] argue that log-normal better represents social networks than a power law. McCormick et. al. [8] also found a log-normal form for the size of a personal network, the total number of people a person knows.

Algorithm Sketch

Let G_a be Alice's graph and let G_b be Bob's graph. Alice and Bob run the following algorithm. They also run the algorithm with their roles reversed and return the best of the two cliques, since Bob may have received the bigger clique piece. If any set is too large to send (super-polylogarithmic), just stop. The adversary gave Alice too few nodes from S. 1) Alice finds the subset of nodes, Q_a with maximum triangle density. Triangle density is the number of triangles divided by the number of nodes. We can find this in polynomial time using a triangle-density version of the edge-density linear program in [2]. 2) Alice finds the set of nodes, $N_a(Q_a)$, adjacent to at least half the nodes in Q_a in the graph G_a and sends Q_a and $N_a(Q_a)$ to Bob. 3) Bob computes $N_b(Q_a)$, the set of nodes adjacent to at least half the nodes Q_a in G_b and sends it to Alice. 4) Let $V_a \leftarrow Q_a \cup N_a(Q_a) \cup N_b(Q_a)$. Let E_a be the set of edges in G_a induced by V_a and let E'_a be the set of edges in G_b induced by V_a. Bob computes E'_a and sends it to Alice. Let $E_a \leftarrow E_a \cup E'_a$. 5) Alice finds the maximum clique in the graph (V_a, E_a) using any intelligent algorithm guaranteed to find the maximum clique.

Correctness Sketch

Let S be the nodes in the planted clique. We can show using Ramsey theory that one of Alice or Bob will receive $\Theta(\ln^3 n)$ triangles of S. We assume without loss of generality that Alice is a player receiving this number of triangles. Any node not in S is involved in $O(1)$ triangles before the clique planting by the clustering-coefficient assumption. Using the maximum-degree assumption, simple probability, the uniform, random selection of clique nodes, and the union bound, we can show that any node not in S has at most a constant number of edges into S with high probability (whp). Therefore any such node has a constant number of triangles involving any node of S. Thus any node not in S is involved in $O(1)$ triangles whp.

Alice's subgraph $Q_a \subseteq S$ whp. The subgraph Q_a has triangle density $\Omega(ln^2 n)$, since Alice received $\Theta(\ln^3 n)$ triangles of the clique with $\ln n$ nodes. In a subgraph of optimal triangle density d, any node participates in $\Omega(d)$ triangles. Otherwise, density would increase by dropping that node. Since any node $v \notin S$ is part of $O(1)$ triangles, it will not be in Q_a. Since Q_a has triangle density $\Omega(\ln^2 n)$, and the maximum triangle density of a graph with x nodes is $O(x^3/x) = O(x^2)$, we have $|Q_a| = \Omega(\ln n)$. In fact, $|Q_a| = \Theta(\ln n)$ because $Q_a \subseteq S$.

The other nodes in S are neighbors of each node in Q_a. Therefore each such node will be adjacent to at least half the nodes in Q_a in G_a and/or G_b. Thus $S \subseteq Q_a \cup N_a(Q_a) \cup N_b(Q_a)$. If there are any stray nodes with high degree into Q_a (a low probability event), the clique-finding operation at the end will remove them. Because $|Q_a| = \Theta(\ln n)$, even exhaustive enumeration runs in polynomial time.

Experiments

In our preliminary experiments, if we find an extremely triangle dense subgraph of size much more than $\ln n$, we remove it as abnormal behavior described by Rossi et. al. [9]. With one such cleaning, we found a planted clique S with $|S| < 3 \ln n$ in the YouTube graph with 3 million edges.

Comments

The structure of social graphs with reasonably-justified restrictions on degree and clustering coefficient allows efficient finding of planted cliques of size $O(\ln n)$. This appears to be much easier than finding cliques in half-dense Erdös-Renyi graphs, where the largest clique is of size $O(\ln n)$, but the best algorithms can only find planted cliques of size $\Theta(\sqrt{n/e})$[4].

References

[1] T. Britton, M. Deijfen, and A. Martin-Löf. Generating simple random graphs with prescribed degree distribution. *Journal of Statistical Physics*, 124(6), September 2006.

[2] M. Charikar. Greedy approximation algorithms for finding dense components in a graph. In *Proceedings of the Third International Workshop on Approximation Algorithms for Combinatorial Optimization*, pages 84–95, 2000.

[3] F. Chung and L. Lu. The average distances in random graphs with given expected degrees. *PNAS*, 99:15879–15882, 2002.

[4] Y. Deshpande and A. Montanari. Finding hidden cliques of size $\sqrt{n/e}$ in nearly linear time. *arxiv*, 1304(7047v1), 2013.

[5] R. Dunbar. Social cognition on the internet: testing constraints on social network size. *Philosophical Transactions of the Royal Society B, Biological Sciences*, 367(1599):2192–2201, 2012.

[6] D. Easley and J. Kleinberg. Networks, crowds, and markets. *Cambridge Univ Press*, 6(1):6–1, 2010.

[7] T. Kolda, A. Pinar, T. Plantenga, and C. Seshadhri. A scalable generative graph model with community structure. *arxiv*, 1302(6636v1), 2013.

[8] T. H. McCormick, M. J. Salganik, and T. Zheng. How many people do you know?: Efficiently estimating personal network size. *Journal of the American Statistical Association*, 105(489), mar 2010.

[9] R. A. Rossi, D. F. Gleich, A. H. Gebremedhin, and M. M. A. Patwary. A fast parallel maximum clique algorithm for large sparse graphs and temporal strong components. *arxiv*, 1302(6256v1), 2013.