DIMACS Technical Report 2006-10
May 2006

Finding max-cliques in power-law graphs with large clique coefficients

(PRELIMINARY VERSION)

by

James Abello and Michael Capalbo[1]

[1]the second author was supported by grant number EIA02-05116

ABSTRACT

Here we show that the Max-Clique problem is hard, even when restricted to sparse low-diameter power-law graphs with a large clustering coefficient.

1 Introduction

The Max-Clique problem is as follows: Given a graph G, find the largest clique that is a subgraph of G. Unfortunately, it has already been shown that it is NP-hard to find an approximation algorithm for the Max-Clique problem on general n-vertex graphs that always returns a clique of size within even a factor of $n^{1-\varepsilon}$ of the optimum, where ε is an arbitrarily small constant. For example, see [1], [2], [3].

In this note, we essentially prove that the Max-Clique problem, even restricted to sparse, low-diameter power-law graphs with a high clustering coefficient, is hard. This question is interesting for the following reason. When studying the cause-and-effect of (even noncontagious) diseases, a patient-attribute graph is often used. The *patient-attribute graph*, or equivalently, the PAG (in a study of a particular disease), is a bipartite graph G where on one side is the set of vertices representing the patients in the study that have acquired the disease, and on the other side is a collection of *attributes* that may be associated with the disease (such as age of the patient, patient eating habits, where the patient lives, etc). There is an edge in the PAG between the vertex representing a patient, and the vertex representing a certain attribute, if the patient has that attribute. For example, if Patient A smokes and smoking is an attribute represented, then there is an edge in the PAG from the vertex representing Patient A to the vertex representing smoking. Now, finding large bicliques in a PAG where both sides are large is of interest to epidemiologists, because this can be used to identify a set S of attributes (and causes) that is prevalent enough among patients to conclude that there is a significant positive correlation between the attributes in S, and the likelyhood of having the disease. This might be quite helpful in targeting prevention.

So to make the connection between finding large balanced cliques in PAG's and finding large cliques in (not necessarily bipartite) power-law graphs, we note the following. Empirical evidence shows that PAG's have a power-law structure. In fact, in the detailed analysis of SEER data (which is a large set of US Cancer data, see `http://seer.cancer.gov/about/`) performed last year by James Abello and David Millman, they noticed that the PAG obtained from SEER data is sparse, of low diameter and with small clique size. They also noted that PAG's are double-covers of sparse, low-diameter power-law graphs. So finding a large clique in a sparse low-diameter power-law graph corresponds to finding a large balanced biclique in a PAG. So with that in mind, we focused on finding either a good algorithm for finding large cliques in power-law graphs, or a certificate that this is impossible unless P = NP.

We now present our notation, and then we present our formal result. Let F be a graph, and let v be a vertex in F. We define $d_F(v)$ to be the degree of v in F. We define $\lambda_F(v)$ to be the number of edges e in F such that *both* endpoints of e are adjacent to v. Finally, we define the quantity $cc_F(v)$ to be the quantity $\frac{\lambda_F(v)}{\binom{d_F(v)}{2}}$, and cc_F to be $\min_{v\in V(F)} cc_F(v)$. So if F is a clique, then cc_F is 1. On the other hand, if F has no cycle with fewer than 4 vertices, then $cc_F(v)$ is 0 for each vertex $v \in V(F)$. Finally, we say that $H = (m_1, \ldots, m_n)$, where the m_i's are nonnegative intgers with m_n positive, is the *degree sequence* of a graph G if G has m_i vertices of degree i, for each integer i, and if $\sum_{i=1}^{n} m_i = |V(G)|$. We are now ready to state the main result.

1.1 The main result

The rest of this paper will focus on the proof of Theorem 1.1, stated below.

Theorem 1.1 *First, let H be any degree sequence for a graph on $\hat{N}$ vertices that satisfies the conditions*

(A) $\sum_{j=\frac{n'}{2}}^{n'} m_j > \frac{n}{2}$ for all n' satisfying $4 \le n' \le n$ (where $m_1 = m_2 = m_3 = 0$),

(B) n is at least N^α, for some $\alpha \in \Omega(1)$,

(C) $\sum_{j=4}^{n} jm_j = O(\hat{N}\log\hat{N})$, and

(D) $\sum_{j=\frac{n}{2}}^{n} jm_j$ is $o(\sum_{j=k}^{n} m_j)$ for each $k \in O(1)$.

Next, let c be any real number no larger than $\frac{1}{27}$. Next, let $\mathcal{Z}_{H,c}$ be the family of graphs on N vertices such that each $Z \in \mathcal{Z}_{H,c}$ satisfies

(I) Z has diameter $O(\log\hat{N})$, and

(II) the quantity cc_Z is at least c.

Then the MAX-CLIQUE problem, even restricted to $\mathcal{Z}_{H,c}$, is NP-hard.

To see how Theorem 1.1 relates to power-law graphs, note that Conditions (A)–(D) are indeed slight relaxations on the restrictions on the degree sequences of sparse power-law graphs; i.e., the degree sequence of every sparse power-law graph satisfies (A)–(D), but not every graph whose degree sequence satisfies (A)–(D) can be considered power-law.

2 Proof of Theorem 1.1

We prove Theorem 1.1 in the following fashion. First, let $\hat{N}$ be a sufficiently large integer, and fix a degree sequence $H = (m_1, \ldots, m_n)$ for a graph on $\hat{N}$ vertices that satisfies (A)–(D), and then fix a $c \le 1/27$. Next, denote as $\mathcal{Z}$ the family $\mathcal{Z}_{H,c}$. Then, let G be an arbitrary connected graph with no more than $\frac{n}{10}$ edges. We construct in polynomial time another graph $\hat{Z} \in \mathcal{Z}$ that contains G as a subgraph, and such that finding any clique K that has size within a constant factor of the maximum-sized clique in Z, implies finding a clique in G that has size within a constant factor of the maximum-sized clique in G. Since H satisfies Condition (B) of Theorem 1.1, this suffices to prove the theorem.

We construct $\hat{Z}$ in 7 stages. In Stage 1 we partition $\hat{V}$ into two sets, V_{low} and V, and then we construct a graph Z_{low} on V_{low}. Then in Stages 2–6 we construct another graph Z on V (by first constructing a graph W, then another graph X from W, and then Y from X, and finally, Z from Y). Then in Stage 7 we join the two constructions so that the resulting graph $\hat{Z}$ is as claimed by Theorem 1.1.

Stage 1 Let $\hat{V}$ be a set of vertices such that, writing $H = (m_1, \ldots, m_n)$, the vertex-set $\hat{V}$ satisfies $|\hat{V}| = \sum_i m_i$. The vertex-set of $\hat{Z}$ is $\hat{V}$. Next, let f be a function on $\hat{V}$ such that, for each integer i, the number of vertices v such that $f(v) = i$ is m_i. (For each $v \in \hat{V}$, the degree of v in $\hat{Z}$ will be $f(v)$.) Then let V_{low} be the set of vertices v such that $f(v) \leq 117$. Then construct a graph Z_{low} that satisfies conditions (A)–(D) of Lemma 2.1, stated below.

Lemma 2.1 *We can construct a graph Z_{low} on V_{low}, in polynomial time, that satisfies the following.*

(A) Each vertex in Z_{low} has degree $f(v)$, except possibly one vertex v, which might have degree $f(v) - 1$.

(B) The diameter of Z_{low} is $O(\log |V_{low}|)$, and

(C) $cc_{Z_{low}} \geq \frac{1}{27}$.

(D) There is a set E' of at least $2n$ edges such that $Z_{low} \setminus E''$ satisfies (A), (B), and (C), for any subset E'' of E', and furthermore, we can find such an E' in polynomial time.

Proof: We construct Z_{low} as follows.

(1) Partition V_{low} into sets $V^2_{low}, V^3_{low}, \ldots$ where V^j_{low} is the set of vertices v such that $f(v)$ satisfies $2^j \leq f(v) \leq 2^{j+1}$.

(2) Next, partition each V^j_{low} into *clumps* of 2^j or $2^j + 1$ vertices (this is possible because 2^j is no larger than 117, and each V^j_{low} has $\frac{n}{2}$ vertices because of Condition A of Theorem 1.1. Then let X_{low} be the graph where there are no edges between clumps, and every clump forms a clique. Each vertex $v \in V^j_{low}$ has degree no larger than 2^j, which is no larger than $f(v)$, and each vertex in a clump of 4 vertices has degree 3, which is smaller than $f(v)$. Also, each vertex in X_{low} has degree at least $\frac{f(v)}{2}$ and clustering coefficient 1.

(3) We form Y_{low} from X_{low}. First, for each clump C on 5 or more vertices, remove 2 of the edges from X_{low} that have both endpoints in C, such that those 2 edges form a matching. Next, let M be the number of clumps, and let Λ be a graph on $\{1, \ldots, M\}$ such that Λ has maximum degree 3 and diamter $O(\log M)$. Then place an edge between Clump i and Clump j if and only if there is an edge in Λ between i and j. The resulting graph is Y_{low}, and each vertex v has degree no larger than $f(v)$. Then Y_{low} has diameter $O(\log |V_{low}|)$, and the clustering coefficient of Y_{low} is at least $2/3$.

(4) Add edges to Y_{low} until every vertex v in the resulting graph Z_{low} has degree $f(v)$ (except perhaps one vertex w, which might have degree $f(w)-1$ instead). More specifically, add edges one by one between nonadjacent vertices u and v that are incident to fewer than $f(u)$ and $f(v)$ vertices, respectively, so that (a) the resulting graph is simple, and (b) the set of vertices v that are incident to fewer than $f(v)$ vertices form a clique. Then do the following.

For each such v, let $d(v)$ be the number of edges incident to v. We say that v is *underdegreed* if v is still fewer than $f(V)$ edges. Because $f(v)$ is no larger than 117, it follows that the sum of the $d(v)$'s is also $O(1)$. Then set $D = \sum_v d(v)$. Now there are at least $\frac{n}{100} > D$ clumps of 50 or more vertices each. So for each of these clumps, remove an edge e. Then add put in two edges e_1 and e_2, where one endpoint of e_1 is an underdegreed vertex, the other endpoint of e_1 is an endpoint of e, and one endpoint of e_2 is an underdegreed vertex, and the other endpoint of e_2 is the other endpoint of e. The resulting graph is Z_{low} and satisfies (A), and it is a tedious but easy calculation to check that Z_{low} also satisfies (B), and that the clustering coefficient of Z_{low} is at least $\frac{1}{12}$.

Now, let E' be any set of edges such that (a) both endpoints of each $e \in E'$ are in a clump with more than 50 vertices, and (b) no vertex is incident to any more than 8 edges in E'. Then because H satisfies (*), it follows that $|E|$ is at least $2n$, and it is a tedious but straightforward calculation to prove that $Z_{low} \setminus E''$ satisfies (A), (B), and (C), for each subset E'' of E', and Lemma 2.1 follows. ■

Stage 2 Having consteucted Z_{low} is Stage 1, we start the construction of Z on the remaining set of vertices $V = \hat{V} \setminus V_{low}$. First, partition V into sets $V^0, V^1, V^2, \ldots,$, where V^j is the set of vertices v such that $f(v)$ satisfies $2^{-(j+1)}n \leq f(v) \leq 2^{-j}n$. Next, for each j, further partition V^j into sets $V_1^j, V_2^j, V_3^j, V_4^j$ of equal cardinality (to within 1) such that $\sum_{v \in V_1^j \cup V_2^j} f(v)$ and $\sum_{v \in V_3^j \cup V_4^j} f(v)$ differ by no more than $2^{-j}n$. Then for each i. set $V_i = \cup_j V_i^j$.

Stage 3 Add edges to V_1^1 such that the resulting graph is isomorphic to G, and call the resulting graph W. (By the assumption that H satisfies (A) in the statement of Theorem

1.1, there are at least $\frac{n}{8}$ vertices in V_1^1, and by the assumption that G has no more than $\frac{n}{10}$ edges and is connected, it follows that this is indeed possible.)

Stage 4 Having constructed W, add edges to W in the following fashion to form X. Now, because H satisfies (A) in the statement of Theorem 1.1, it follows that V_i^j has at least $\frac{n}{8}$ vertices. So, for each i and j, and for some integer l_j, partition V_i^j further into sets $V_i^{j,1}, V_i^{j,2} \ldots,$, such that (a) each V_i^{j,l_j} has somewhere between $2^{-(j+4)}n$ and $2^{-(j+3)}n$ vertices, and (b) $V_i^{j,l} = V_{i'}^{j,l}$ for all $i, i' \in \{1,2,3,4\}$, and for all but at most one $l \in \{1,\ldots,l_j\}$. Then, for each l and j, connect every vertex in $V_i^{j_l}$ with every vertex in $V_{i'}^{j,l}$ if and only if i and i' are distinct. The resulting graph is X.

Stage 5 Having constructed X, add edges to X to form Y. Let Λ be any bipartite max-degree 3 graph (with parts $V_1 \cup V_2$ and $V_3 \cup V_4$) and with diameter $O(\log N)$. Then let Y be the graph on V where $E(Y) = E(X) \cup E(\Lambda)$.

Stage 6 Having constructed Y, construct Z from Y in the following fashion. Recall that ultimately, each vertex v in V will have degree $f(v)$ in $\hat{Z}$. Each vertex in V^j has degree at most $3 + 3 \times 2^{-(j+3)}n + D_W$ in Y, where D_W is the maximum degree of W, and that $f(v)$ is at least $\max\{2^{-(j+1)}n, 117\}$. Because the only vertices in V incident to an edge in W are those in V^0, and because W has no more than $|E(G)| \leq \frac{n}{10}$ edges, it follows that the degree in Y of each vertex $v \in V$ is strictly smaller than the degree of each vertex in $\hat{Z}$. So for each vertex $v \in V$, set $t(v) = f(v) - d_Y(v)$.

First, if $\sum_{v \in V_1 \cup V_2} t(v)$ is smaller than $\sum_{v \in V_3 \cup V_4} t(v)$, set $Q = V_1 \cup V_2$; otherwise, set $Q = V_3 \cup V_4$. Then, using the max-flow algorithm in the proof of Lemma 2.2 stated below, add edges to Y between $V_1 \cup V_2$, and $V_3 \cup V_4$ such that (a) each vertex $v \in Q$ is now incident to exactly $f(v)$ edges, and (b) no vertex $v \in V$ is incident to no more than $f(v)$ edges, and (c) the resulting graph is simple. Call the resulting graph Z.

Lemma 2.2 *Let Q be as above. There is a polynomial-time algorithm to add edges to Y between $V_1 \cup V_2$, and $V_3 \cup V_4$ such that (a) each vertex $v \in Q$ is now incident to exactly $f(v)$ edges, (b) the resulting graph is still simple.*

Proof: Set up the following network flow problem $\mathcal{F}$ (with underlying directed graph J). Each vertex v in $V_1 \cup V_2$ has demand $t(v)$, each vertex $y \in V_3 \cup V_4$ has capacity $t(y)$, and there is a directed edge (with capacity 1) in J from $v \in V_1 \cup V_2$ to $y \in V_3 \cup V_4$ if and only if v and y are NOT adjacent in Y. We next show that there exists a flow F with value $\min\{\sum_{y \in V_1 \cup V_2} t(y), \sum_{y \in V_3 \cup V_4} t(y)\}$. However, it is well-known that the existence of such a flow implies that we can find such a flow F in polynomial time such that F is also integral (each directed edge from $V_1 \cup V_2$ to $V_3 \cup V_4$ either has the full one unit of flow or no flow at all). Furthermore, any such flow that is integral will correspond to a set of edges to add to Y such that the resulting graph Z is simple, and satisfies the degree restrainsts promised in Theorem 1.1. So to prove Lemma 2.2, it remains to show existence of such a flow F.

First, for every subset U in $V_1 \cup V_2$, and any vertex $y \in V_3 \cup V_4$, let us write as $g_U(y)$ the quantity $\min\{t(y), r_U(y)\}$, where $r_U(y)$ is the number of vertices in U that are adjacent in J to y, and let us write g_U as the quantity $\sum_{y \in V_3 \cup V_4} g_U(y)$. To show existence of such a flow F, it suffices to show the following (by Menger's Theorem): $g_U \geq -A + \sum_{v \in U} t(v)$ for any set $U \in V_1 \cup V_2$, where $A = |\sum_{y \in V_1 \cup V_2} t(y) - \sum_{y \in V_3 \cup V_4} t(y)|$. We do this now.

Now, observe that g_U is at least $|V_3 \cup V_4| - n$ for every $U \subseteq V_1 \cup V_2$, because every vertex in J has degree at least $|V_3 \cup V_4| - n$. Also note (from Condition (D) in the statment of Theorem 1.1) that there is no set U of $O(n)$ vertices such that $\sum_{v \in U} t(v)$ is $\Omega(|V_3 \cup V_4|)$. So if $|U|$ is no larger than $2n$, then g_U is larger than $\sum_{v \in U} t(v)$. So we are done with this case. On the other hand, if U is larger than $2n$, then because every vertex v in Y has degree no larger than n, it follows that $r_U(y)$ must be at least n for each $y \in V_3 \cup V_4$, which is larger than $t(y)$. So g_U is at least $\sum_{y \in V_3 \cup V_4} t(y)$. So we are done with this case as well. And so by Menger's Theorem, Lemma 2.2 follows. ■

Stage 7 Having constructed Z, we next specify how to add and delete edges from $Z \cup Z_{low}$ so that the resulting graph $\hat{Z}$ is as claimed in the statement of Theorem 1.1.

However, because of how we partitioned the vertices between V_1, V_2, V_3, V_4 in Stage 0, and how we partitioned the V_i^j's in Stage 4 (in particular, condition (b)), it follows that $\sum_{v \in V_1 \cup V_2} g(v)$ and $\sum_{v \in V_3 \cup V_4} g(v)$ are within an additive term η from each other that is no larger than $2n$. So first, call the vertices v in V such that the number of edges incident to v is less than $f(v)$ *underdegreed*. Then identify E'. Then, until there are no more underdegreed vertices, remove an $e \in E'$, and then add 2 edges e_1 and e_2, the first endpoint of each e_i incident to an underdegreed vertex, and the other endpoint of e_1 incident to one endpoint of e, and the other endpoint of e_2 incident to the other endpoint of e. Then when there are no more underdegreed vertices, the resulting graph is $\hat{Z}$.

So the resulting graph $\hat{Z}$ indeed has H as a degree sequence. Also, $\hat{Z}$ has diamter $O(\log \hat{N})$. So all we need to do is to lower bound $cc_{\hat{Z}}$ (via Lemma 2.3 and then show that a large clique in $\hat{Z}$ corresponds to a large clique in G, and we are done. We next prove Lemma 2.3.

Lemma 2.3 *The resulting graph $\hat{Z}$ satisfies $cc_{\hat{Z}} \geq \frac{1}{27}$.*

Proof: From Lemma 2.1, we note that $cc_{\hat{Z}}(v)$ is at least $\frac{1}{27}$ for each $v \in V_{low}$. So to finish the proof, we just need to show that $cc_{\hat{Z}}(v)$ is at least $\frac{1}{27}$ for each vertex $v \in V$. We do this next.

Consider Stage 4. Each vertex v such that $f(v)$ satisfies $2^{-(j+1)}n \leq f(v) \leq 2^{-j}n$ has a subset S_v of at least $3 \times 2^{-(j+4)}n$ vertices $y \in V$ that are neighbors in X, where each such y in turn has $2 \times 2^{-(j+4)}n$ neighbors in S_v. So $\lambda_{\hat{Z}}(v)$ is at least $6 \times 2^{-2(j+4)}n^2$ (because no edges incident to a vertex in V were taken away after Stage 4). But the degree of v in $\hat{Z}$ is $f(v)$, which is no larger than $2^{-j}n$, and so Lemma 2.3 follows. ∎

From Lemma 2.3 we see that $\hat{Z}$ satisfies (II) of Theorem 1.1. So to finish the proof of Theorem 1.1, it suffices to prove Lemma 2.4.

Lemma 2.4 *Let K be any clique in $\hat{Z}$. Then we can find in time polynomial in $|V(G)|$ a clique in G of size at least $|V(K)| - 121$ vertices.*

Proof: For each i, consider $K \cap V_i$, and then consider $K \cap V_{low}$. The induced subgraph of $\hat{Z}$ on each V_1 is isomorphic to G, and is empty on V_2, V_3, and V_4. So there may be at most one vertex from each V_i; $i = 2, 3, 4$, in any clique in $\hat{Z}$. Also, the induced subgraph of $\hat{Z}$ on V_{low} has maximum degree 117. So there may be at most 118 vertices from V_{low} in any clique in $\hat{Z}$. So Lemma 2.4 follows. ∎

So Theorem 1.1 follows from Lemma 2.4. ∎

References

[1] L. Engebretsen and J. Hastad. Clique is hard to approximate within $n^{1-o(1)}$. In *Proceedings of 27th International Colloquium on Automata, Languages and Programming.* Volume 1853, LNCS, pages 2–12, 2000.

[2] J. Hastad. Clique is hard to approximate within $n^{1-\epsilon}$. *Acta Mathematica*, 182:105–142, 1999.

[3] S. Knot. Improved inapproximability results for max-clique, chromatic number and approximate graph coloring. In *Proceedings of 42nd Annual IEEE Symposium on Foundations of Computer Science*, pp 600–609, Las Vegas, 2001.

[4] http://seer.cancer.gov/about/.