Regular graphs whose subgraphs tend to be acyclic

Noga Alon [*]   Eitan Bachmat [†]

Abstract

Motivated by a problem that arises in the study of mirrored storage systems, we describe, for any fixed $\epsilon, \delta > 0$ and any integer $d \geq 2$, explicit or randomized constructions of d-regular graphs on $n > n_0(\epsilon, \delta)$ vertices in which a random subgraph obtained by retaining each edge, randomly and independently, with probability $\rho = \frac{1-\epsilon}{d-1}$, is acyclic with probability at least $1 - \delta$. On the other hand we show that for any d-regular graph G on $n > n_1(\epsilon, \delta)$ vertices, a random subgraph of G obtained by retaining each edge, randomly and independently, with probability $\rho = \frac{1+\epsilon}{d-1}$, does contain a cycle with probability at least $1-\delta$. The proofs combine probabilistic and combinatorial arguments, with number theoretic techniques.

1 Introduction

The moment of appearance of the first cycle in an evolving random graph has been studied extensively in [5]. It is known that the first cyclic component appears on average when the graph has approximately $0.44n$ edges where n is the number of vertices, however, this random variable has a huge variance, and there is a positive probability of containing a cycle when there are only εn edges, for any $\varepsilon > 0$.

Random graphs may be viewed as random subgraphs of the family of complete graphs. In this paper we consider the appearance of the first cycle in the evolution of a random subgraph of certain families of d-regular graphs, where d is fixed. In particular we are interested in constructing such families in which the appearance of the first cycle is postponed as much as possible.

As will be explained later on, this problem was originally motivated by the problem of designing efficient configurations for mirrored storage systems. The results here also yield an interesting application in Coding Theory.

Let G be a graph. Our model for picking a random edge subset A will be the standard random subgraph model in which each edge of G is chosen independently with some fixed probability ρ. We let $G(\rho)$ denote the probability space on the subgraphs of G induced by this model.

[*] Schools of Mathematics and Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel. Email: noga@math.tau.ac.il. Research supported in part by a USA-Israeli BSF grant, by the Israel Science Foundation and by the Hermann Minkowski Minerva Center for Geometry at Tel Aviv University.

[†] Department of Computer Science, Ben-Gurion University, Email:ebachmat@cs.bgu.ac.il

Consider a family $\bar{G} = (G_n)$ of d-regular graphs. Consider the random variable $\chi_{n,\rho} : G_n(\rho) \to \{0, 1\}$ whose value on A is 1 iff the subgraph of G induced by the edge set A is acyclic.

We define the *cycle threshold* of the family $\bar{G}$ as

$$CT(\bar{G}) = Sup \left\{ \frac{1}{2} d\rho \mid \liminf_{n\to\infty} E(\chi_{n,\rho}) = 1 \right\}$$

Note that $\frac{1}{2}d\rho|V(G_n)|$ is the expected number of edges in a subgraph in $G_n(\rho)$, and therefore $CT(\bar{G})$ represents the asymptotic ratio of edges to vertices for the maximum ρ for which the first cycle is expected not to occur with high probability.

In this paper we study families of regular graphs with large cycle thresholds. Our new results, and the organization of the rest of the paper, are as follows. In Section 2 We prove that $CT(\bar{G}) \leq \frac{d}{2(d-1)}$ for any family of d-regular graphs. In Section 3 we show that the families of graphs of number theoretic origin constructed by Lubotzky, Phillips and Sarnak [9] achieve this upper bound for each $d = p + 1$, with $p \equiv 1(\ Mod \ 4)$ a prime. In Section 4 we show that by modifying appropriately random d-regular graphs we obtain, with high probability, regular graphs which asymptotically achieve the upper bound as well. Section 5 contains the description of an application to storage systems configuration design, and another application in Coding Theory.

1.1 Some definitions and Notation

A *cycle* in G is a connected 2- regular subgraph of G. The length of a cycle C, denoted $l(C)$, is the number of its edges. A (non backtracking) *walk* W in G from a vertex v to a vertex u is a sequence of vertices $v = v_0, v_1, \ldots, v_m = u$ in G such that $(v_i, v_{i+1}) \in E(G)$ and $v_i \neq v_{i+2}$ for all $i = 0, ..., m-2$. The length $l(W)$ of such a walk is m. We note that if C is a cycle and v is a vertex of C, then C induces two walks of length $l(C)$ from v to itself by going around the cycle in either direction, hence the number of walks of length k from v to itself is at least twice the number of cycles of length k containing v. The girth of a graph G, denoted $g(G)$, is the minimum cycle length in G. Throughout the paper, the degree of regularity d is considered to be fixed, whereas the size n of the graphs grows to infinity. All logarithms are in the natural base e, unless otherwise specified.

2 An upper bound on $CT(\bar{G})$

In this section we prove the following result. The proof combines the approach of [1] with some additional ideas.

Theorem 2.1 *Let $\bar{G}$ be a family of d-regular graphs, where $d \geq 2$. Then $CT(\bar{G}) \leq \frac{d}{2d-2}$. Moreover, the following stronger result holds. For any integer $d \geq 2$ and any real $\epsilon > 0$, there is a finite $n_0 = n_0(,\epsilon)$ so that the following holds. For any $n > n_0$ and any d-regular graph G on n vertices, if $\rho = \frac{1+\epsilon}{d-1}$, then the probability that the random subgraph $G(\rho)$ of G, obtained by retaining each edge, randomly and independently with probability ρ, is acyclic, does not exceed ϵ.*

Proof. Fix $d \geq 2$ and $\epsilon > 0$, and let $G = (V, E)$ be a d-regular graph on n vertices. Throughout the proof we assume, whenever this is need, that n is sufficiently large as a function of d and ϵ.

Whenever we write a term $o(1)$, we refer to a quantity that tends to zero, as n tends to infinity.

Define $s = \lfloor \log_d \log n \rfloor$, and let S be a maximal (with respect to containment) collection of pairwise edge disjoint cycles of length at most s in G. Note, first, that if $|S| \geq \log^2 n$, then, as each cycle $C \in S$ lies completely in $G(\rho)$ with probability $\rho^{|C|} \geq \rho^s > \frac{1}{\log n}$, the probability that $G(\rho)$ is acyclic is at most the probability it contains no member of S which is bounded by

$$(1 - \frac{1}{\log n})^{|S|} < \frac{1}{n} < \epsilon,$$

as needed. Thus we may and will assume that $|S| \leq \log^2 n$. We claim that there is a set $X \subset V$ of at least, say, $\sqrt{n}$ vertices of G, satisfying the following.
(i) The distance between any vertex of X and any cycle in S is at least $4s$.
(ii) The distance between any two vertices in X is at least $4s$.

Indeed, the total number of vertices that lie within distance $4s$ of some cycle in S is bounded by

$$|S|sd^{4s} < O(\log^7 n).$$

We can thus pick the vertices of X one by one, always choosing a vertex that does not lie within distance $4s$ of any of the members of S, and does not lie within distance $4s$ of any of the previously chosen members of X. As long as we have chosen less than $\sqrt{n}$ vertices of X, there are still less than

$$O(\log^7 n) + \sqrt{n}d^{4s} \leq O(\sqrt{n} \log^4 n)$$

vertices that cannot be picked, providing the required set X of size at least $\sqrt{n}$ (with room to spare).

By the choice of X and the maximality of S, each vertex $v \in X$ does not lie within distance s of any cycle of length at most s in G (since each such cycle intersects at least one member of S, and this member is far from v.) Therefore, seen from any vertex $v \in X$, the graph G out to a distance s looks just like a d-regular tree.

It is convenient to consider the random subgraph $G(\rho)$ as a union of two independently chosen random subgraphs $G(\rho_1)$ and $G(\rho_2)$, where $\rho_1 = \frac{1+\epsilon/2}{d-1}$ and ρ_2 ($\geq \frac{\epsilon}{2d}$) is chosen such that $(1 - \rho_1)(1 - \rho_2) = 1 - \rho$. Pick a vertex $v \in X$ and expose all the edges of $G(\rho_1)$ that lie (in G) within distance at most s of v. By standard results from the theory of branching processes (see, for example, [7]), there is an $\eta > 0$ such that with probability at least η the connected component of v in this exposed subgraph of $G(\rho_1)$ has at least, say, $(1 + \epsilon/10)^s$ leaves. We next show that conditioning on this event, the random graph $G(\rho)$ will contain a cycle with probability $1 - o(1)$. Indeed, conditioning on this event, we keep following this branching process. To do so, choose in each step a yet unexplored leaf u of the connected component, and in case there are $d - 1$ edges of G emanating from it to vertices outside this component, expose the corresponding $d - 1$ edges of $G(\rho_1)$. If this is not the case, that is, if u is adjacent in G to at least one other vertex of the component besides its parent in the component, we declare u to be an explored vertex, and keep it as a leaf. Note that whenever we scan the edges of an unexplored vertex, the behavior is precisely as in the usual branching process on the infinite d-regular tree, and hence, since $\rho > \frac{1}{d-1}$ and the number of leaves in the component grows with n, the component will keep growing with high probability, and will stop with a component K containing among its leaves at least $(1+\Omega(1))^s$ leaves, each having at least one neighbor in K besides its parent. We can now expose the random edges of $G(\rho_2)$, and if these edges contain any one of the edges connecting such a leaf to its neighbor in K besides its parent, we get a cycle in $G(\rho)$. As the number of these edges grows with n, this happens with probability $1-o(1)$, and we conclude that indeed if the component of v in $G(\rho_1)$ grows at the beginning, as assumed, then a cycle emerges with high probability.

In case the component of v in $G(\rho_1)$ dies early, we pick another vertex $v' \in X$ and repeat the same process from there. Since each such vertex gives a growing component with probability at least η, and we have $\sqrt{n}$ vertices, we conclude that with probability $1-o(1)$ there will be a vertex $w \in X$ whose component will grow, providing, with high probability, the required cycle. This completes the proof. ■

Remark: In the interesting case of families of high girth, regular expander graphs, much more precise information on the components of $G_n(\rho)$ for $\rho > 1/(d-1)$ can be found in [1].

3 Explicit families with $CT(\bar{G}) = \frac{d}{2d-2}$

In this section we prove the existence of explicit families of d-regular graphs with $CT(\bar{G}) = \frac{d}{2d-2}$, using well known number theoretic graph constructions

The LPS construction

We first describe a well known family of graphs, the bipartite LPS Ramanujan graphs, constructed by Lubotzky, Philips and Sarnak in [9]. We follow closely the exposition in the first three sections of [9]. Let p, q be primes satisfying $p, q = 1$ $(Mod\ 4)$. Let F_q be the field with q elements and let $i \in F_q$ satisfy $i^2 = -1$ $(Mod\ q)$, (it is easy and well known that such an i exists, as $q = 1$ $(Mod\ 4)$.)

Define the quadratic residue symbol $(\frac{p}{q})$ to be 1 if the equation $x^2 = p(Mod\ q)$ is solvable in F_q, and -1 otherwise. It follows from quadratic reciprocity and the well known theorem of Dirichlet on primes in arithmetic progressions, that for each fixed prime $p > 2$ there are infinitely many primes q with $q = 1(Mod\ 4)$ and $\frac{p}{q} = -1$. It is also known by a formula of Jacobi that the equation

$$x_0^2 + x_1^2 + x_2^2 + x_3^2 = p \tag{1}$$

has $p+1$ solutions with $x_0 > 0$ and x_1, x_2, x_3 even. Let $PGL_2(q)$ denote the group of invertible 2 by 2 matrices with elements in F_q, where we identify two matrices if they are proportional to each other. Assume that $a = (a_0, a_1, a_2, a_3)$ is a solution to the equation (1) above, and let $\gamma(a)$ be the matrix

$$\begin{pmatrix} a_0 + a_1 i & a_2 + a_3 i \\ -a_2 + a_3 i & a_0 - a_1 i \end{pmatrix} \tag{2}$$

considered as an element in $PGL_2(q)$. Consider the Cayley graph of $PGL_2(q)$ with respect to the symmetric set $\{\gamma(a)\}$ of size $p+1$. We define $X_{p,q}$ to be the connected component of the identity matrix in this graph. Obviously $X_{p,q}$ is vertex-transitive. We also have $|V(X_{p,q})| \leq |PGL_2(q)| = q^3 - q \leq q^3$.

Consider the set of integer Hamiltonian quaternions $a_0 + a_1 i + a_2 j + a_3 k$, with $a_i \in \mathbf{Z}$, where multiplication is given by the relations $i^2 = j^2 = k^2 = -1$ and $ij = -ji = k$, $jk = -kj = i$, $ki = -ik = j$. For an element $\alpha = a_0 + a_1 i + a_2 j + a_3 k$, define the conjugate $\bar{\alpha} = a_0 - a_1 i - a_2 j - a_3 k$ and the norm

$$N(\alpha) = \alpha\bar{\alpha} = a_0^2 + a_1^2 + a_2^2 + a_3^2.$$

Given p let $\Lambda'(2)$ be the set of quaternions with a_0 odd, a_1, a_2, a_3 even and with $N(\alpha) = p^k$ for some integer k. $\Lambda'(2)$ is closed under multiplication. We identify elements $\alpha, \beta \in \Lambda'(2)$ whenever $\alpha = (-1)^i p^j \beta$ for some i, j, and denote the group of resulting classes $[\alpha]$ by $\Lambda(2)$. Given this equivalence relation, any element in $\Lambda(2)$ can be represented as a class $[\alpha]$ for a unique $\alpha \in \Lambda'(2)$ satisfying $a_0 > 0$ odd, a_1, a_2, a_3 even and not all a_i divisible by p. We shall call such an α a normalized element. Note also that $[\alpha]^{-1} = [\bar{\alpha}]$. Let $\alpha_m = a_{0,m} + a_{1,m} i + a_{2,m} j + a_{3,m} k$, $m = 1, \ldots, p+1$, be the set of quaternions whose coefficients form the $p+1$ solutions of equation (1). We assume that the α_m are indexed so that $\alpha_{m+(p+1)/2} = \bar{\alpha}_m$ for $m = 1, ..., (p+1)/2$. Using Jacobi's formula and unique factorization in the Hamilton quaternions, it is shown in [9] (Lemma 3.1 and Corollary 3.2) that the group $\Lambda(2)$ is freely generated by $[\alpha_1], ..., [\alpha_{(p+1)/2}]$. Thus, the corresponding Cayley graph is the infinite d-regular tree. Moreover $X_{p,q}$ is the corresponding Cayley graph of the quotient of $\Lambda(2)$ by the normal subgroup Γ_q defined by the elements $\alpha \in \Lambda'(2)$ for which a_1, a_2, a_3 are all divisible by $2q$. (See [9] for more details.)

Walks of length k in the graph $X_{p,q}$, beginning and ending at the identity element, correspond to paths in the universal tree cover, starting at the identity and ending at an element of Γ_q. By the discussion above such paths correspond in turn to products $[\alpha_{i_1}] \cdot ... \cdot [\alpha_{i_k}] \in \Gamma_q$. The assumption that no backtracking is allowed in a walk and the freeness of the generators $[\alpha_i]$ means that $\alpha = \alpha_{i_1} \cdot ... \cdot \alpha_{i_k}$ is normalized up to its sign. Since the norm on the quaternions is multiplicative, we have $N(\alpha) = p^k$. Let W_v denote the number of such walks in $X_{p,q}$ which start and end at a vertex v. Since $X_{p,q}$ is transitive we conclude that for each v, W_v is equal to the number of integer solutions of the equation

$$x_0^2 + q^2(x_1^2 + x_2^2 + x_3^2) = p^k \tag{3}$$

where $x_0 > 0$ is odd, x_1, x_2, x_3 are even, and at least one of x_0, x_1, x_2, x_3 is not divisible by p. We call such solutions normalized solutions. This fact, which is implicit in [9], is the key to our computations.

Computation of $CT(\bar{G})$

We now present families of LPS graphs with optimal $CT(\bar{G})$. The proof generalizes the girth lower bound argument for LPS graphs, given in [9]. For the sake of completeness we reprove the theorem with a slightly stronger result than that stated in [9].

Theorem 3.1 *([9] theorem 3.4) Let p, q be primes as in the LPS construction. Assume that $(\frac{p}{q}) = -1$ and assume there exists a walk of size k in $X_{p,q}$ corresponding to a normalized solution x_0, x_1, x_2, x_3 of (3), then the following hold.*

1) All possible values of x_0 lie in the union of two arithmetic progressions with difference $2q^2$.

2) $p^k \geq q^4$

3) $g(X_{p,q}) \geq (4/3) \log_p |V(X_{p,q})|$

Proof. Consider a normalized solution of (3). Since

$$x_0^2 = p^k \ (Mod\ q^2) \tag{4}$$

and p is not a quadratic residue modulo q, k must be even. It is well known that $\mathbf{Z}^*_{\mathbf{q^2}}$ is cyclic, thus equation (4) has only two solutions, which in this case must be $p^{k/2}\ (Mod\ q^2)$ and $-p^{k/2}\ (Mod\ q^2)$. Combining this with the fact that x_0 is odd we obtain part 1. The solution $x_0 = p^{k/2}, x_1 = 0, x_2 = 0, x_3 = 0$ to equation (3) is not normalized, thus if $p^{k/2} \leq q^2$ there is no normalized solution with $x_0 = p^{k/2}(\ Mod\ q^2)$, as the other possible (positive) values of $x_0 = p^{k/2}(\ Mod\ q^2)$ are too large to satisfy (3). Therefore we must have $x_0 = q^2 - p^{k/2}$ which is not possible since x_0 must be odd. Hence, $p^{k/2} > q^2$, establishing part 2. Part 3 follows from part 2 and the inequality $|V(X_{p,q})| \leq q^3$. ■

We can now prove the main result of the section.

Theorem 3.2 *Fix a prime $p = 1\ (Mod\ 4)$, put $d = p + 1$ and let q_n be an increasing sequence of primes all equal to 1 modulo 4 and satisfying $(\frac{p}{q_n}) = -1$. Let $\bar{G}$ be the family of graphs $G_n = X_{p,q_n}$, then*

$$CT(\bar{G}) = \frac{d}{2d-2}.$$

Proof. Let $c_{k,n}$ denote the number of cycles of length k in G_n. Let $r_{k,n} = r_{p,q_n}(k)$ be the number of normalized integer solutions to equation (3). Since the numbers of cycles of size k is bounded by the number of closed walks of size k we have

$$c_{k,n} \leq r_{k,n}|V(G_n)| \leq r_{k,n} q_n^3.$$

Any x_0 in a solution to equation (3) must satisfy $|x_0^2| \leq p^k$ hence $|x_0| \leq p^{k/2}$. By part 1 of Theorem 3.1 any such x_0 must belong to a union of two arithmetic progressions with difference $2q^2$. Thus there are at most $O(p^{k/2}/q_n^2)$ choices for x_0. Given one of these choices of x_0 we consider $m = (p^k - x_0^2)/q_n^2 = x_1^2 + x_2^2 + x_3^2$. Obviously $|m| \leq p^k/q_n^2$. We also have $x_1^2 \leq m$. As x_1 can be either positive or negative but must be even, for each fixed x_0 there are at most $\sqrt{m} \leq p^{k/2}/q_n$ possible choices for x_1. Having chosen x_1 as well we need to consider all solutions to $x_2^2 + x_3^2 = m - x_1^2 \leq p^k$. By theorem 338 of [6], for any $\varepsilon > 0$ there is a constant A'_ε such that there are at most $A'_\varepsilon p^{\varepsilon k}$ solutions to this equation. We thus have

$$r_{p,q_n}(k) \leq O(p^{k/2} q_n^{-2})(p^{k/2} q_n^{-1})(A'_\varepsilon p^{\varepsilon k}) \leq A_\varepsilon p^{(1+\varepsilon)k}/q_n^3 \tag{5}$$

solutions to equation (3). We deduce that $c_{k,n} \leq A_\varepsilon p^{(1+\varepsilon)k}$. Suppose, now, that the edges of the graph G_n are chosen randomly and independently with probability ρ, forming the random subgraph $G_n(\rho)$. Then the probability of a given cycle of size k to be chosen is ρ^k. Let $\rho = p^{-(1+\delta)}$ for some $\delta > 0$, and let ε be such that $\delta > \varepsilon > 0$. Let $C_{n,\rho}$ be the random variable which counts the number of cycles in $G_n(\rho)$. By the first moment method $E(C_{n,\rho})$ is an upper bound on the probability that a graph in the $G_n(\rho)$ model contains a cycle. Let $V_n = |V(X_{p,q_n})|$, then

$$E(C_{n,\rho}) = \sum_{k=3}^{\infty} c_{k,n}\rho^k = \sum_{(4/3)\log_p(V_n)}^{\infty} c_{k,n}\rho^k$$

$$\leq \sum_{(4/3)\log_p(V_n)}^{\infty} A_\varepsilon(p^{\varepsilon-\delta})^k = O(V_n^{(4/3)(\varepsilon-\delta)}).$$

The last quantity tends to 0 for every fixed $\delta > \varepsilon > 0$, implying the desired result. ■

The Morgenstern construction

The LPS construction produced families of d-regular graphs when $d - 1 \equiv \ (1 \ Mod \ 4)$ is a prime. In [12] Morgenstern generalized this construction to give families of $q + 1$-regular graphs where q is any prime power. The construction is analogous to the LPS construction. The graphs obtained exhibit similar properties and enable us to prove the following result, whose detailed proof is omitted.

Theorem 3.3 *For any odd prime power q there is an infinite family $\bar{G}$ of $d = q + 1$-regular graphs, whose cycle threshold satisfies $CT(\bar{G}) = \frac{q+1}{2q}$.* ■

It is interesting to note that the proof of Theorem 3.2 does not apply to non-bipartite Ramanujan LPS graphs of size n, whose girth is only known to be at least $\frac{2}{3}\log_{d-1} n$. In fact, it can be shown that any family of vertex transitive, d-regular graphs with a nearly optimal cycle threshold, must have bigger girth. This is proved in the following proposition.

Proposition 3.4 *For any $d > 2$ and $\epsilon > 0$ the following holds. For any d-regular vertex transitive graph G with n vertices, whose girth does not exceed $(1-\epsilon)\log_{d-1} n$, the random subgraph obtained by retaining each edge of G, randomly and independently, with probability $\rho = \frac{1}{(d-1)^{1+\epsilon}}$ contains a cycle with probability at least*

$$1 - e^{-n^{\epsilon^2}/\log_{d-1}^2 n}$$

Proof. Fix a cycle C of minimum length $s \leq (1-\epsilon)\log_{d-1} n$ in G. We claim that G contains a family $\mathcal{F}$ of at least $\frac{n}{s^2}$ pairwise vertex disjoint cycles of length s. Indeed, let $\mathcal{F}$ be a family of such cycles of maximum cardinality. Assume the claim is false, and $|\mathcal{F}| < \frac{n}{s^2}$. Let f be a random automorphism of the graph, chosen uniformly among all automorphisms, and consider the cycle $C' = f(C)$. Obviously C' is of length s. In addition, since for each vertex v of C, $f(v)$ is distributed uniformly among all vertices of G, the probability that $f(v)$ belongs to one of the cycles in $\mathcal{F}$ is precisely $|\mathcal{F}|s/n < 1/s$. It follows that with positive probability, $C' = f(C)$ does not intersect any member of $\mathcal{F}$, contradicting the maximality and proving the claim.

Suppose, now, that $G(\rho)$ is a random subgraph of G obtained by retaining each edge, randomly and independently, with probability $\rho = \frac{1}{(d-1)^{1+\epsilon}}$. As the cycles in $\mathcal{F}$ are pairwise edge-disjoint, the probability that $G(\rho)$ does not contain any of them is precisely

$$(1-\rho^s)^{|\mathcal{F}|} \le \left(1-(\frac{1}{(d-1)^{1+\epsilon}})^{(1-\epsilon)\log_{d-1} n}\right)^{n/s^2} = (1-\frac{1}{n^{1-\epsilon^2}})^{n/s^2} \le e^{-n^{\epsilon^2}/\log^2_{d-1} n}$$

This completes the proof. ■

4 Families of random graphs

The families of graphs described in the previous section have asymptotically optimal cycle thresholds. However these families exist only if $d = q + 1$ where q is a prime power. Moreover, if we denote the number of vertices of each graph in a family by n, then even for degrees d as above the construction yields graphs only for a sparse set of values of n. In applications (to be described in the next section) one would like to construct graphs whose subgraphs tend to be acyclic for all admissible values of d and n.

Looking at the proof of Theorem 3.2 we notice that there are two properties which allowed us to prove optimality. The first is high girth and the second is that for any k there are at most $O((d-1)^{k(1+\varepsilon)})$ cycles of size k in the graph. Random d-regular graphs have the second property but not the first. There are several ways to generate high-girth random or pseudo-random d-regular graphs of given size n, see, for example, [4], [11], but it seems difficult to prove that they retain the second property. It may be possible to apply the techniques of [11] to show that with positive (though exponentially small) probability, a random d-regular graph on n vertices has girth at least $\Omega(\log_{d-1} n)$ and does not contain more than $O((d-1)^k)$ cycles of length k, for any k. Since, however, it seems unlikely that this method will lead to an efficient algorithm for generating such graphs, we prefer to apply a different approach. We thus consider random d-regular graphs on n labeled vertices using the well studied configuration model (see, e.g., [3], [8]), and then modify them in order to eliminate the few short cycles they contain, keeping the property of having a relatively small number of longer cycles.

Motivated by the discussion above we describe a simple, efficient, randomized procedure to generate, for every fixed integer d, and fixed real $\epsilon > 0$, and for any large integer n so that nd is even, a d-regular graph G on n vertices, such that a random subgraph of G obtained by keeping each edge of G, randomly and independently, with probability $\frac{1-\epsilon}{d-1}$, is acyclic with high probability.

The precise statement of the result is the following.

Theorem 4.1 *Let $d \ge 2$ be a fixed integer, and let $\epsilon > 0$, $\delta > 0$ be positive reals. Define $k_0 = k_0(\epsilon, \delta) = \lceil \frac{1}{\epsilon} \ln(\frac{1}{\delta}) \rceil$ and $K = K(d, \epsilon, \delta) = 16d^{k_0+1}$. Then, for every integer $n > K$ so that nd is even, there exists a d-regular graph G on n vertices, such that a random subgraph of G obtained by keeping each edge of G, randomly and independently, with probability $\rho = \frac{1-\epsilon}{d-1}$, is acyclic with probability at least $1 - \delta - \frac{2K}{\epsilon} n^{-\epsilon/2\log d}$. Moreover, there is an efficient randomized algorithm that produces a graph that has this property with probability at least $1/4$.*

Remark: The constants above can be improved, we make no attempt to optimize them. Both ϵ and δ in the theorem may be functions of n. In particular, by taking them to be some functions that tend slowly to zero as n grows, e.g., $\epsilon = \delta = 1/\log\log\log n$, we get a family of d-regular graphs with optimal cycle threshold $\frac{d}{2d-2}$. Although these graphs may well have girth much smaller than $\log_{d-1} n$, this does not contradict the assertion of Proposition 3.4, as the graphs constructed in the proof below are not vertex transitive.

Proof. Fix d, ϵ, δ and a sufficiently large n so that nd is even. Assume that $d \geq 3$, as the result for $d < 3$ is trivial. Let $H = (V, E)$ be a random d-regular graph on n labeled vertices generated according to the configuration model described in [3], [8]. This is done by taking a uniform random perfect matching on the set $W = \{v_{ij} : \ 1 \leq i \leq n, 1 \leq j \leq d\}$ and then by collapsing each group $V_i = \{v_{ij}, 1 \leq j \leq d\}$ to a vertex v_i. The number of configurations in this model, that is, the number of perfect matchings on W, is $(nd-1)!! = (nd-1)(nd-3)(nd-5)\cdots 1$. For every integer k satisfying $1 \leq k \leq n$, the number of these configurations leading to a graph in which the vertices $v_1, v_2, \ldots, v_k$ form a simple cycle of length k in this order is at most

$$2(\binom{d}{2})^k 2^{k-1}(nd-2k-1)!!$$

As there are $\frac{1}{2k}n(n-1)(n-2)\cdots(n-k+1)$ potential simple cycles of length n on the vertices $v_1, v_2, \ldots, v_n$, this implies that the expected number of simple cycles of length k in H is at most

$$c(n,d,k) = \frac{1}{2k}\frac{n(n-1)\cdots(n-k+1)(d(d-1))^k}{(nd-1)(nd-3)\cdots(nd-2k+1)}.$$

This number is at most $\frac{1}{2k}(d-1)^k$ for all $k \geq 3$, and at most $\frac{1}{2k}(d-1)^k(1+O(\frac{1}{nd}))$ for $k = 1, 2$.

Let C_k denote the number of cycles of length k in H, and let $\rho = \frac{1-\epsilon}{d-1}$ be as in the theorem. By linearity of expectation, the expected number of cycles of length at most k_0 in H is at most $\sum_{k=1}^{k_0} c(n,d,k) < 2d^{k_0+1} = K/8$, where K is as defined in the statement of the theorem. Therefore, with probability at least $7/8$, the number of such cycles is at most K.

Put $k_1 = \frac{1}{2}\log_d n \ (\ > k_0)$. As in the previous paragraph, the expected number of cycles of length at most k_1 in H is at most $\sum_{k=1}^{k_1} c(n,d,k) < 2d^{k_1+1} < \frac{1}{8}n^{2/3}$, and hence with probability at least $7/8$, the number of such cycles is at most $n^{2/3}$.

Let $X = X(H)$ be the random variable defined as follows: $X = \sum_{k>k_0} C_k \rho^k$. Again, linearity of expectation gives that the expectation of X is at most

$$\sum_{k>k_0} \frac{1}{2k}(d-1)^k \rho^k < \frac{1}{2k_0}\sum_{k>k_0}(1-\epsilon)^k \leq \frac{\delta}{2}.$$

Therefore, with probability at least $1/2$, the value of X is at most δ. It follows that with probability at least $1/4$, H has at most K cycles of length at most k_0, at most $n^{2/3}$ cycles of length at most k_1, and the random variable X computed at H is at most δ. Fix such a graph H. The desired graph G will be obtained from H by performing at most K *switching operations*, as described below, in order to destroy all cycles of length at most k_0, without creating any new cycles of length at most k_1, and without creating too many longer cycles.

Let $e_1 = \{u_1, v_1\}$, $e_2 = \{u_2, v_2\}$ be two edges in a graph H', where in each edge e_i, u_i is considered the first vertex, and where $e_1' = \{u_1, u_2\}$, $e_2' = \{v_1, v_2\}$ are non-edges. The graph obtained from H' by switching e_1, e_2 is the graph H'' obtained from H' by deleting the edges e_1, e_2 and by adding the two new edges e_1', e_2'. Note that if H' is d-regular, then so is H''. We claim that if the distance in H between the two edges e_1, e_2 is at least k_1, and e_2 lies in no cycle of length at most k_1, then the switching operation creates no new cycles of length at most k_1. Indeed, if a new cycle contains only one of the newly added edges, say $\{u_1, u_2\}$, then it must contain a path in H' from u_1 to u_2, and by assumption, the length of any such path is at least k_1. If a new cycle contains both newly added edges, then it must contain either a path in H' from u_2 to v_2, or a path in H' from u_2 to v_1, and in both cases the resulting cycle is of length exceeding k_1. Another simple observation is the fact that for any k and for any edge e in a d-regular graph, e can lie in less than $(d-1)^k$ cycles of length k (as the number of walks of length $k-1$ starting at a vertex is bounded by $(d-1)^{k-1}$.) Since any switching operation adds two new edges, it can add at most $2(d-1)^k$ new cycles of length k, for any k.

Returning to our graph H, we now modify it to obtain the desired graph G as follows. Starting with H, as long as our graph contains a cycle of length at most k_0, pick an arbitrary edge e_1 in it, and pick another edge e_2 of distance at least k_1 from e_1 which does not lie on a cycle of length at most k_1. Then switch e_1, e_2. As this process creates no new cycles of length at most k_1, throughout the process our graph contains at most $n^{2/3}$ cycles of length at most k_1. Therefore, at most $n^{2/3}k_1 < n^{2/3}\log n$ edges lie on such cycles, and as the number of edges within distance k_1 from e_1 is at most $2d^{k_1} = 2\sqrt{n}$, there is always a valid choice for e_2. Each such switching operation destroys the cycle of length at most k_0 through e_1, and hence this process must terminate after at most K steps. By the discussion above, this gives a graph G of girth exceeding k_0, in which the number of cycles of each length $k \le k_1$ is precisely C_k- the number of cycles of that length in H. Moreover, the number of cycles of length k in G for larger values of k is at most $C_k + 2K(d-1)^k$.

Suppose, now, that $G(\rho)$ is a random subgraph of G obtained by picking each edge of G, randomly and independently, with probability ρ. Then the expected number of simple cycles in $G(\rho)$ is at most

$$\sum_{k=k_0+1}^{k_1} C_k\rho^k + \sum_{k>k_1}(C_k + 2K(d-1)^k)\rho^k = x(H) + \sum_{k>k_1} 2K(d-1)^k\rho^k$$

$$= x(H) + 2K\sum_{k>k_1}(1-\epsilon)^k \le \delta + \frac{2K}{\epsilon}n^{-\epsilon/2\log d}.$$

It follows that the probability that $G(\rho)$ contains a cycle does not exceed $\delta + \frac{2K}{\epsilon}n^{-\epsilon/2\log d}$, as needed.

The randomized algorithm to generate G is simple: generate H, find all its cycles of length at most k_1 (by checking all walks of that length), and then perform the switchings as in the proof. This completes the proof. ∎

5 Applications

In this section we briefly survey the intended application of the results above to the problem of configuring mirrored storage devices. We first consider a concrete example, configuring an online video store, and then explain the more general context of the application. A more thorough examination of these issues will be presented in [2]. In addition, we describe an interesting application in Coding Theory.

5.1 Online video stores and mirrored storage systems

Assume we want to create an online video store. We would like to lower our costs as much as possible. Since the customers do not take the videos home, we can place several movies together on the same high capacity DVD. The only problem is that a DVD can only show one movie at a time. Let n be the number of DVDs. We assume that we can place d movies on each DVD. If we only have a single copy of each movie, then the number of randomly chosen movies which we will be able to display concurrently without refusing a client is $O(\sqrt{n})$; this is simply the birthday paradox. We are also unprotected against device failures. If we allow two copies of each movie, then we can describe the layout of the movies on the storage devices (DVDs or tapes) by a *configuration graph* G. The vertices of the graph correspond to the storage devices, while the edges of the graph correspond to the movies. Each movie connects the two devices on which it resides. Let A be the subset consisting of all movies the customers wish to view concurrently at a given time slot. We can show all the movies if and only if we can find a one-to-one assignment map $f_{G,A} : A \longrightarrow V(G)$ such that $f(e)$ is incident to e for all $e \in A$. The assignment function tells us which of the two storage devices containing copies of the movie corresponding to an edge e will be used to show it.

Storage devices sometimes fail, and we would like to be able to show the required set of movies A even after the failure of any single storage device in the system (multiple, simultaneous device failures are very rare). This can be done if for any vertex $v \in G$ we can find a one-to-one assignment mapping $f_{G,A,v} : A \longrightarrow V(G) - \{v\}$, that is, an assignment of A which does not use the storage device corresponding to v to show any movie.

If assignment functions $f_{G,A,v}$ as above exist for all v, we say that the configuration G supports A *exclusively* since each device is responsible for the screening of a single movie, thus the movie receives exclusive service. The following simple observation directly relates the problem of exclusive support to the problem studied in the previous sections

Observation ([13], [14]) : A can be supported exclusively by G if and only if it is acyclic.

In view of this observation it is clear that the cycle threshold of $\bar{G}$ provides an estimate for the number of movie requests the store can expect to support asymptotically. Theorem 2.1 tells us that we cannot hope to support more than $\frac{d}{2(d-1)}n$ requests with our n DVDs. Theorems 3.2 and 4.1 show how to construct asymptotically optimal systems which achieve the $\frac{d}{2(d-1)}n$ bound (assuming the requests are random).

More generally, almost all storage devices in use today, such as disks, DVDs and tapes, are mechanical devices. The service time of such devices heavily depends on the amount of time the device spends transitioning from one data location to another. One method of improving the service time of such devices is to partition the storage space of the device into d local regions with relatively small internal transition times and to attempt to restrict each device to service exclusively only one region. In fact this is the only method that can be used to control maximal service times at the device level itself (other methods such as caching are aimed at reducing device traffic). We assume that the data of each region is mirrored on another region on a different device. The results of the paper state that the strategy is likely to succeed if at any given moment most of the activity is concentrated on at most $\frac{d}{2(d-1)}n$ of the $\frac{dn}{2}$ data regions (assumed to be chosen randomly). This places some constraints on our ability to deliver high quality of service requirements using mechanical storage devices. A more thorough examination of these issues will be given in [2], together with experimental results that indicate that random subgraphs of bipartite LPS Ramanujan graphs tend to have a higher probability of being acyclic than random subgraphs of random graphs with the same parameters n, d.

Remark: One may consider the more general notion of $k - exclusive$ service in which each device serves at most k regions. This notion is obviously important for studying load balancing properties of mirrored storage systems. This notion (under a different name) has been studied extensively in [13] and [14] using some properties of random graphs (i.e., random subgraphs of complete graphs).

5.2 Cycle Codes

High girth graphs with a relatively small number of cycles of any length, like the bipartite Ramanujan LPS graphs described in Section 3, can be used to construct linear, binary error-correcting codes. Although these codes are not as good asymptotically in terms of minimum distance and rate as are some more complicated constructions, they exhibit some appealing properties. Here is a brief description. For more background and basic properties of error correcting codes the reader is referred to [10].

Given an undirected graph $G = (V, E)$, let $C = C(G)$ denote the linear code consisting of all binary vectors $f(e)_{e \in E}$ satisfying $\sum_{e; v \in e} f(e) = 0$ for all $v \in V$, where addition is computed modulo 2. Therefore, C is simply the cycle space of the graph G, consisting of all characteristic vectors of subgraphs of G in which all degrees are even. If G is connected and has m edges and n vertices, then the dimension of this code is $m - n + 1$, and its length is m. Its minimum distance, that is, the minimum weight of a codeword, is the girth of G, which, for any graph with $m \geq (1 + \epsilon)n$ for some fixed ϵ, is at most $c(\epsilon) \log n$. Although this is much weaker than the linear distance that can be achieved by more sophisticated constructions, these codes have some nice properties. The first such property is the fact that encoding is extremely simple: pick an arbitrary spanning tree T in the graph, put the message bits on all non-tree edges, and use the tree-edges as parity check bits to make sure that the resulting word is a codeword. This can be easily done by scanning the vertices of the tree from leaves to root, where in each vertex v in its turn, the value of the bit on the edge from v to its parent is chosen so that the sum of all bits on the edges incident with v is even.

Another appealing property of these codes is the fact that maximum likelihood decoding can also be performed here efficiently. That is, given any word $g = g(e)_{e \in E}$, one can find efficiently the codeword $f(e)_{e \in E}$ which is closest to g. Indeed, viewing g as the characteristic vector of a subgraph H of G, let U be the set of all vertices of odd degree in H. The objective is to find a collection $E' \subset E$ of minimum cardinality, so that the degree of a vertex in the graph (V, E') is odd iff the vertex lies in U. The required closest codeword f is simply the sum modulo 2 of the characteristic vector of E' with g. Finding E' can be done by applying any minimum weight matching algorithm; this is known as the T-join problem, see, e.g., [15], pp. 486-487 for more details. Note that in fact such an optimal decoding can be performed efficiently even if we have a probability for the identity of the bit on each edge, and the objective is to find the codeword which maximizes the product of the individual resulting probabilities.

Finally, suppose that starting with a codeword $f(e)_{e \in E}$, the bit on every edge is flipped, randomly and independently, with probability ρ. This gives a word g, and our decoding procedure enables us to find the codeword f' which is closest to g. What is the probability of error, that is the probability that $f' \neq f$? This probability can be bounded as follows. The sum modulo 2 of f and f' is another codeword, that is, a member of the cycle space of G, and hence it is the characteristic vector of a union of pairwise edge-disjoint cycles of G. If f' is closer to g than f, then there is at least one cycle of G in which more than half the bits on its edges have been flipped. If we let c_k denote the number of cycles of length k in our graph, and we let k_0 denote its girth, then this probability does not exceed

$$\sum_{k \geq k_0} c_k B_k$$

where

$$B_k = \sum_{i > k/2} \binom{k}{i} \rho^i (1-\rho)^{k-i}$$

is the probability that a binomial random variable with parameters k and ρ exceeds $k/2$. The graphs discussed in Section 3 have large girth and asymptotically the smallest possible values for the numbers c_k among all d-regular graphs. They thus provide codes in which the probability of error is polynomially small in n, provided $c_k B_k \leq (1-\epsilon)^k$ for all k and some fixed ϵ. In particular, if we use the 4-regular Ramanujan graphs of girth at least $\frac{4}{3}\log_3 n$, constructed in [12], and take, say $\rho \leq 1/36$, we get a code of length $2n$, dimension $n+1$, with efficient encoding and (maximum likelihood) decoding schemes, and with probability of error which is polynomially small in n.

References

[1] N. Alon, I. Benjamini and A. Stacey, Percolation on finite graphs and isoperimetric inequalities, Ann. Probab., in press.

[2] N. Alon, E. Bachmat and U. Shapenko, Mirrored storage configurations for supporting QoS requirements, in preparation.

[3] B. Bollobás, A probabilistic proof of an asymptotic formula for the number of labeled regular graphs, European J. Combinatorics 1 (1980), 311-316.

[4] B. Bollobás, *Extremal Graph Theory*, Academic press, 1978.

[5] P. Flajolet, D.E. Knuth and B. Pittel, The first cycles in an evolving graph, Discrete Mathematics 75 (1989), 167-217.

[6] G.H. Hardy and E.M.Wright, *Introduction to Number Theory*, Forth Edition, Oxford U. Press, 1975.

[7] T. Harris, *The Theory of Branching Processes,* Springer-Verlag, Berlin (1963).

[8] S. Janson, T. Łuczak and A. Ruciński, *Random Graphs*, Wiley, New York, 2000.

[9] A. Lubotzky A., R. Phillips and P. Sarnak, Ramanujan graphs, Combinatorica, 8 (1988), 261-277.

[10] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*, North Holland, Amsterdam, 1977.

[11] B. D. McKay, N. W. Wormald and B. Wysocka, Short cycles in random regular graphs, to appear.

[12] M. Morgenstern., Existence and explicit construction of $q+1$ regular Ramanujan graphs for every prime power q, Journal Combinatorial Theory, Series B 62 (1994), 44-62.

[13] P. Sanders, Reconciling simplicity and realism in parallel disk models, Parallel Computing, 28 (2002), 705-723.

[14] P. Sanders, S. Egner and J. Korst, Fast concurrent access to parallel disk, *Proc. of the 11th SODA conference*, 849-858, 2000.

[15] A. Schrijver, *Combinatorial Optimization*, Springer-Verlag, 2003.