Dynamic concentration of the triangle-free process

Tom Bohman[1] and Peter Keevash[2]

Abstract. The triangle-free process begins with an empty graph on n vertices and iteratively adds edges chosen uniformly at random subject to the constraint that no triangle is formed. We determine the asymptotic number of edges in the maximal triangle-free graph at which the triangle-free process terminates. We also bound the independence number of this graph, which gives an improved lower bound on Ramsey numbers: we show $R(3,t) > (1/4 - o(1))t^2/\log t$, which is within a $4 + o(1)$ factor of the best known upper bound. Furthermore, we determine which bounded size subgraphs are likely to appear in the maximal triangle-free graph produced by the triangle-free process: they are precisely those triangle-free graphs with maximal average density at most 2.

1 Introduction

Constrained random graph processes provide an interesting class of random graph models and a natural source for constructions in graph theory. Although the dependencies introduced by the constraints make such processes difficult to analyse, the evidence to date suggests that they are particularly useful for producing graphs of interest for certain extremal problems. Here we consider the triangle-free random graph process, which is defined by sequentially adding edges, starting with the empty graph, chosen uniformly at random subject to the constraint that no triangle is formed.

This process was introduced by Bollobás and Erdős (see [7]), and first analysed by Erdős, Suen and Winkler [10], using a differential equations method introduced by Ruciński and Wormald [17] for the analysis of the constrained graph process known as the 'd-process'. One motivation for their work was that their analysis of the triangle-free process led to the best lower bound on the Ramsey number $R(3,t)$ known at that time. The Ramsey number $R(s,t)$ is the least number n such that any graph on

[1] Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, USA. Email: tbohman@math.cmu.edu. Research supported in part by NSF grants DMS-1001638 and DMS-1100215.

[2] School of Mathematical Sciences, Queen Mary, University of London, Mile End Road, London E1 4NS, UK. Email: p.keevash@qmul.ac.uk. Research supported in part by ERC grant 239696 and EPSRC grant EP/G056730/1.

J. Nešetřil et al. (eds.), *The Seventh European Conference on Combinatorics, Graph Theory and Applications*

n vertices contains a complete graph with s vertices or an independent set with t vertices. In general, very little is known about these numbers, even approximately. The upper bound $R(3,t) = O(t^2/\log t)$ was obtained by Ajtai, Komlós and Szemerédi [1], but for many years the best known lower bound, due to Erdős [9], was $\Omega(t^2/\log^2 t)$. The order of magnitude was finally determined by Kim [13], who showed that $R(3,t) = \Omega(t^2/\log t)$. He employed a semi-random construction that is loosely related to the triangle-free process, thus leaving open the question of whether the triangle-free process itself achieves this bound; this was conjectured by Spencer [19] and proved by Bohman [3]. There is now a large literature on the general H-free process, obtained by replacing 'triangle' by any fixed graph H in the definition; see [6,8,15,16,22–25]. However, the theory is still in its early stages: we conjectured that our lower bound for H strictly 2-balanced, given in [6], gives the correct order of magnitude for the length of the process, but so far this has only been proved for some special graphs.

In this paper we specialise to the triangle-free process, where we can now give an asymptotically optimal analysis. Our improvement on previous analyses of this process exploits the self-correcting nature of key statistics of the process.

Let G be the maximal triangle-free graph at which the triangle-free process terminates.

Theorem 1.1. *Whp every vertex of G has degree $(1+o(1))\sqrt{\frac{1}{2}n\log n}$.*

We also obtain the following bound on the size of any independent set in G.

Theorem 1.2. *Whp G has independence number at most $(1+o(1))\sqrt{2n\log n}$.*

An immediate consequence is the following new lower bound on Ramsey numbers. The best known upper bound is $R(3,t) < (1+o(1))\frac{t^2}{\log t}$, due to Shearer [18].

Theorem 1.3. $R(3,t) > \left(\frac{1}{4} - o(1)\right)\frac{t^2}{\log t}$.

These results are predicted by a simple heuristic: the graph $G(i)$ after i steps of the triangle-free process should resemble the Erdős-Rényi random graph $G_{n,p}$ with $i = n^2p/2$, with the exception that $G_{n,p}$ has many triangles while $G(i)$ has none. We also show that this heuristic extends to all small subgraph counts; in particular, we answer the question of which subgraphs appear in G. Suppose H is a graph with at least 3 vertices.

The *average density* of H is $d(H) = \frac{|E_H|}{|V_H|}$. The *maximum average density* $m(H)$ of H is the maximum of $d(H')$ over nonempty subgraphs H' of H.

Theorem 1.4. *Let H be a triangle-free graph with at least 3 vertices.*

(i) *If $m(H) \le 2$ then $\mathbb{P}(H \subseteq G) = 1 - o(1)$.*
(ii) *If $m(H) > 2$ then $\mathbb{P}(H \subseteq G) = o(1)$.*

Thus, the small subgraphs that are likely to appear in G are exactly the same as the triangle-free subgraphs that appear in the corresponding $G_{n,p}$.

2 Overview of the proof

We are guided by the heuristic that $G(i)$ resembles $G_{n,p}$ with $i = n^2p/2$. We introduce a continuous time that scales as $t = in^{-3/2}$. Note that $p = 2tn^{-1/2}$. We define $Q(i)$ to be the number of open *ordered* pairs in $G(i)$. This variable is crucial to our understanding of the process: we have $Q(0) = n^2 - n$, and the process ends when $Q(i) = 0$. How do we expect $Q(i)$ to evolve? If $G(i)$ resembles $G_{n,p}$ then for any pair uv we should have $\mathbb{P}(uv \in O(i)) \approx \left(1 - p^2\right)^{n-2} \approx e^{-np^2} = e^{-4t^2}$. We set $q(t) = e^{-4t^2}n^2$ and expect to have $Q(i) \approx q(t)$ for most of the evolution of the process. This is exactly what we prove.

2.1 Strategy

We use dynamic concentration inequalities for a carefully chosen ensemble of random variables associated with the process. We show $V(i) \approx v(t)$ for all variables V in the ensemble, for some smooth function $v(t)$, which we refer to as the *scaling* of V. Here $V(i)$ denotes the value of V after i steps of the process. For each V we define a *tracking variable* $\mathcal{T}V(i)$ and show that $\mathcal{D}V(i) = V(i) - \mathcal{T}V(i)$ satisfies $|\mathcal{D}V(i)| < e_V(t)v(t)$, for some error functions $e_V(t)$. We use $\mathcal{T}V(i)$ rather than $v(t)$ so that we can isolate variations in V from variations in other variables that have an impact on V.

The improvement to earlier analysis of the process comes from 'self-correction', *i.e.* the mean-reverting properties of the system of variables. We take $e_V(t) = f_V(t) + 2g_V(t)$, where we think of $f_V(t)$ as the 'main error term' and $g_V(t)$ as the 'martingale deviation term'. We usually have $g_V \ll f_V$, but there are some exceptions when t is small and hence $f_V(t)$ is too small. We require $g_V(t)v(t)$ to be 'approximately non-increasing' in t, in that $g_V(t')v(t') = O(g_V(t)v(t))$ for all $t' \ge t$. We define the *critical window* $W_V(i) = [(f_V(t) + g_V(t))v(t), (f_V(t) + 2g_V(t))v(t)]$.

We prove the *trend hypothesis*: $\mathcal{Z}V(i) := |\mathcal{D}V(i)| - e_V(t)v(t)$ is a supermartingale when $|\mathcal{D}V(i)| \in W_V(i)$. The trend hypothesis will follow from the *variation equation* for $e_V(t)$, which balances the changes in $\mathcal{D}V(i)$ and $e_V(t)v(t)$. Since errors can transfer from one variable to another, each variation equation is a differential inequality that can involve many of the error functions.

We track the process up to the time $t_{max} = \frac{1}{2}\sqrt{(1/2-\varepsilon)\log n}$. If the tracking fails, then there is some $i^* \le i_{max}$ and a variable V such that $\mathcal{D}V(i)$ enters $W_V(i')$ from below at some step $i' < i^*$, stays in $W_V(i)$ for $i' \le i \le i^*$ then goes above $W_V(i^*)$ at step i^*. During this time $\mathcal{Z}V(i)$ is a supermartingale, with $\mathcal{Z}V(i') \le -g_V(t')v(t')$ and $\mathcal{Z}V(i^*) \ge 0$, so we have an increase of at least $g_V(t')v(t')$ against the drift of the supermartingale. We can estimate the probability of this event using Freedman's martingale inequality [11], provided that we have good estimates on $Var_V(t) = Var(\mathcal{Z}V(i) \mid \mathcal{F}_{i-1})$ and $N_V(t) = |\mathcal{Z}V(i+1) - \mathcal{Z}V(i)|$; we refer to this as the *boundedness hypothesis*. Thus it suffices to verify the trend and boundedness hypotheses for all variables.

2.2 Variables

All definitions are with respect to the graph $G(i)$. Sometimes we use a variable name to also denote the set that it counts, *e.g.* $Q(i)$ is the number of ordered open pairs, and also denotes the set of ordered open pairs. We usually omit (i) and (t) from our notation, *e.g.* Q means $Q(i)$ and q means $q(t)$. We use capital letters for variable names and the corresponding lower case letter for the scaling. We express scalings using the (approximate) edge density and open pair density, namely $p = 2in^{-2} = 2tn^{-1/2}$ and $\hat{q} = e^{-4t^2}$.

The next most important variable in our analysis, after the variable Q defined above, is the variable Y_{uv} which, for a fixed pair of vertices uv, is the number of vertices w such that uw is an open pair and vw is an edge. It is natural that Y_{uv} should play an important role in this analysis, as when the pair uv is added as an edge, the number of open edges that become closed is exactly $Y_{uv} + Y_{vu}$. The motivation for introducing the ensembles of variables defined below is as follows: control of the global variables is needed to get good control of Q, control of the stacking variables is needed to get good control of Y_{uv}, and controllable variables play a crucial role in our analysis of the stacking variables.

The *global variables* consist of Q, R and S, where $Q = 2|O(i)|$ is the number of ordered open pairs, R is the number of ordered triples with 3 open pairs, and S is the number of ordered triples abc where ab is an edge and ac, bc are open pairs.

The *stacking variables* are built from four basic building blocks: X_u is the number of vertices ω such that $u\omega$ is open, Y_u is the number of vertices ω such that $u\omega$ is an edge, X_{uv} is the number of vertices w such that uw and vw are open pairs, Y_{uv} is the number of vertices w such that uw is an open pair and vw is an edge. We defer the exact definition to the full version of the paper, but roughly speaking, the idea is that the relative errors in these variables decrease as the number of steps increases, so that after a large constant number of steps they are essentially global.

Finally, we formulate a very general condition under which we have some control on a variable. Suppose Γ is a graph, J is a spanning subgraph of Γ and $A \subseteq V_\Gamma$. We refer to (A, J, Γ) as an *extension*. Suppose that $\phi : A \to [n]$ is an injective mapping. We define the *extension variables* $X_{\phi,J,\Gamma}(i)$ to be the number of injective maps $f : V_\Gamma \to [n]$ such that f restricts to ϕ on A, $f(e) \in E(i)$ for every $e \in E_J$ not contained in A, and $f(e) \in O(i)$ for every $e \in E_\Gamma \setminus E_J$ not contained in A. We introduce the abbreviations $V = X_{\phi,J,\Gamma}$, $n(V) = |V_\Gamma| - |A|$, $e(V) = e_J - e_{J[A]}$, and $o(V) = (e_\Gamma - e_J) - (e_{\Gamma[A]} - e_{J[A]})$. The *scaling* is $v = x_{A,J,\Gamma} = n^{n(V)} p^{e(V)} \hat{q}^{o(V)}$. We expect $V \approx v$, provided there is no subextension that is 'sparse', in that it has scaling much smaller than 1. Given $A \subseteq B \subseteq B' \subseteq V_\Gamma$ we define $S_B^{B'} = S_B^{B'}(J, \Gamma)$ to equal

$$n^{|B'|-|B|} p^{e_{J[B']} - e_{J[B]}} \hat{q}^{(e_{\Gamma[B']} - e_{J[B']}) - (e_{\Gamma[B]} - e_{J[B]})}.$$

Let $t' \geq 1$. We say that V is *controllable at time* t' if $J \neq \Gamma$ (*i.e.* at least one pair is open) and for $1 \leq t \leq t'$ and $A \subsetneq B \subseteq V_\Gamma$ we have $S_A^B(J, \Gamma) \geq n^\delta$, where $\delta > 0$ is a fixed global parameter that is sufficiently small given ε. (This condition is essentially identical to the condition needed to prove concentration of subgraphs counts in $G_{n,p}$ using Kim-Vu polynomial concentration [14].)

3 Concluding remarks

We have determined $R(3, t)$ to within a factor of $4 + o(1)$, so we should perhaps hazard a guess for its asymptotics: we are tempted to believe the construction rather than the bound, *i.e.* that $R(3, t) \sim t^2/4 \log t$. We should note that we only have an upper bound on the independence number of the graph G produced by the triangle-free process. So, formally speaking, the triangle-free process could produce a graph that gives a better lower bound on $R(3, t)$. But we believe that this is not the case; that is, we conjecture that the bound on the independence number in Theorem 1.2 is asymptotically best possible.

Our method for establishing self-correction builds on ideas used recently by Bohman, Frieze and Lubetzky [5] for an analysis of the triangle-removal process (see also [4] for a simpler context). Furthermore, the

results of this paper have also been obtained independently and simultaneously by Fiz Pontiveros, Griffiths and Morris; their proof also exploits self-correction, but is different to ours in some important ways.

Another natural direction for future research is to provide an asymptotically optimal analysis in greater generality for the H-free process. No doubt the technical challenges will be formidable, given the difficulties that arise in the case of triangles. But on an optimistic note, it is encouraging that one can build on two different proofs of this case.

References

[1] M. AJTAI, J. KOMLÓS and E. SZEMERÉDI, *A note on Ramsey numbers*, J. Combin. Theory Ser. A **29** (1980), 354–360.

[2] N. ALON and J. SPENCER, "The Probabilistic Method", second edition, Wiley, New York, 2000.

[3] T. BOHMAN, *The triangle-free process*, Adv. Math. **221** (2009), 1653–1677.

[4] T. BOHMAN, A. FRIEZE and E. LUBETZKY, *A note on the random greedy triangle-packing algorithm*, J. Combinatorics **1** (2010), 477–488.

[5] T. BOHMAN, A. FRIEZE and E. LUBETZKY, *Random triangle removal*, arXiv:1203.4223.

[6] T. BOHMAN and P. KEEVASH, *The early evolution of the H-free process*, Invent. Math. **181** (2010), 291–336.

[7] B. BOLLOBÁS and O. RIORDAN, *Random graphs and branching processes*, In: "Handbook of Large-scale Random Networks", Bolyai Soc. Math. Stud. **18**, Springer, Berlin, 2009, 15–115.

[8] B. BOLLOBÁS and O. RIORDAN, *Constrained graph processes*, Electronic J. Combin. **7** (2000), R18.

[9] P. ERDŐS, *Graph theory and probability, II*, Canad. J. Math. **13** (1961), 346–352.

[10] P. ERDŐS, S. SUEN and P. WINKLER, *On the size of a random maximal graph*, Random Structures Algorithms **6** (1995), 309–318.

[11] D. A. FREEDMAN, *On tail probabilities for martingales*, Ann. Probability **3** (1975), 100–118.

[12] S. GERKE and T. MAKAI, *No dense subgraphs appear in the triangle-free graph process*, Electron. J. Combin. **18** (2011), R168.

[13] J. H. KIM, *The Ramsey number $R(3, t)$ has order of magnitude $t^2/\log t$*, Random Structures Algorithms **7** (1995), 173–207.

[14] J. H. KIM and V. H. VU, *Concentration of multivariate polynomials and its applications*, Combinatorica **20** (2000) 417–434.

[15] D. OSTHUS and A. TARAZ, *Random maximal H-free graphs*, Random Structures Algorithms **18** (2001), 61–82.
[16] M. PICOLLELLI, *The final size of the C_4-free process*, Combin. Probab. Comput. **20** (2011), 939–955.
[17] A. RUCIŃSKI and N. WORMALD, *Random graph processes with degree restrictions*, Combin. Probab. Comput. **1** (1992), 169–180.
[18] J. SHEARER, *A note on the independence number of triangle-free graphs*, Disc. Math. **46** (1983), 83–87.
[19] J. SPENCER, *Maximal trianglefree graphs and Ramsey $R(3, k)$*, unpublished manuscript.
[20] J. SPENCER, *Asymptotic lower bounds for Ramsey functions*, Disc. Math. **20** (1997), 69–76.
[21] J. SPENCER, *Counting extensions*, J. Combin. Theory Ser. A **55** (1990), 247–255.
[22] L. WARNKE, *Dense subgraphs in the H-free process*, Disc. Math. **333** (2011), 2703–2707.
[23] L. WARNKE, *When does the K_4-free process stop?* Random Structures Algorithms, to appear.
[24] G. WOLFOVITZ, *Lower bounds for the size of random maximal H-free graphs*, Electronic J. Combin. **16**, 2009, R4.
[25] G. WOLFOVITZ, *Triangle-free subgraphs in the triangle-free process*, Random Structures Algorithms **39** (2011), 539–543.