Maximizing the expected number of components in an online search of a graph

Fabrício Siqueira Benevides*   Małgorzata Sulkowska**

*Universidade Federal do Ceará, Brazil
**Wrocław University of Science and Technology, Poland

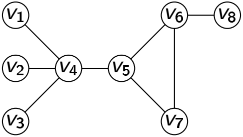
8th Polish Combinatorial Conference
Będlewo, September 2020

Problem formulation

- A selector knows graph $G = (V, E)$, $|V| = N$.
- $\mathcal{S}$ is the set of all permutations of V. Select uniformly at random a permutation $\sigma \in \mathcal{S}$, say $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_N)$.
- Vertices of G are revealed one by one, following the order given by σ. $\tilde{G}_t$ is the graph induced by $\{\sigma_1, \sigma_2, \ldots, \sigma_t\}$.
- At time t (i.e., after t vertices have been revealed) the selector takes a decision based on some information he receives about $\tilde{G}_t$: either he continues the process and reveals the next vertex or he stops the game and gains as payoff the number of connected components of $\tilde{G}_t$ (denoted by $\tilde{C}_t$). His aim is to maximize the expected number of components in the graph $\tilde{G}_t$.
- Three versions of the game are considered depending on the information the selector gets about $\tilde{G}_t$:
  1. blind game,
  2. partial information game,
  3. full information game.

1. **Blind game**
   At time t the selector knows only the number of vertices that have already appeared (i.e., t). He has no other information about the revealed structure. In fact, he gains no information during the game.
2. **Partial information game**
   The selector can see an unlabeled graph isomorphic to $\tilde{G}_t$. In particular, he knows how many edges or components are there at time t, but he does not know exactly which vertices of G have been selected. This is a classical setup for many optimal stopping problems considered in the past (compare setup of the **secretary problem**).
3. **Full information game**
   The selector knows $\{\sigma_1, \sigma_2, \ldots, \sigma_t\}$, and since he knows G, he knows $\tilde{G}_t$. Thus he gets all information that is available at time t.

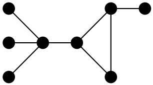
Below we present a course of the game for the same graph G and the same permutation of its vertices σ in three versions: blind, partial information and full information. Consider a graph G:



and permutation $\sigma = (v_7, v_8, v_6, v_2, v_3, v_1, v_4, v_5)$. The selector always knows G in advance and never knows σ in advance.

1. **Blind game, example**

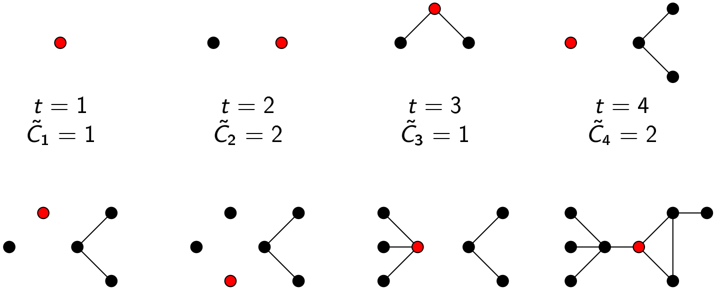
   The selector gets no information during the game. He can actually decide before the game when to stop. He knows states at $t = 1$ and $t = N$ which are always the same.



$t = 1, \tilde{C}_1 = 1$ $\cdots$ $t = 8, \tilde{C}_8 = 1$

2 Partial information game, example

At time t the selector can see an unlabeled graph isomorphic to $\tilde{G}_t$. In particular, he knows $\tilde{C}_t$.



$t = 1$
$\tilde{C}_1 = 1$

$t = 2$
$\tilde{C}_2 = 2$

$t = 3$
$\tilde{C}_3 = 1$

$t = 4$
$\tilde{C}_4 = 2$

$t = 5$
$\tilde{C}_5 = 3$

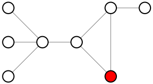
$t = 6$
$\tilde{C}_6 = 4$

$t = 7$
$\tilde{C}_7 = 2$

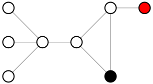
$t = 8$
$\tilde{C}_8 = 1$

3. **Full information game, example**

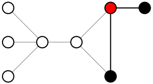
   The selector knows exactly, which vertex of G appears at time t.



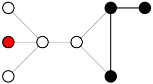
$t = 1$
$\tilde{C}_1 = 1$



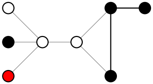
$t = 2$
$\tilde{C}_2 = 2$



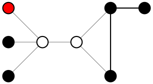
$t = 3$
$\tilde{C}_3 = 1$



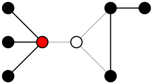
$t = 4$
$\tilde{C}_4 = 2$



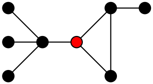
$t = 5$
$\tilde{C}_5 = 3$



$t = 6$
$\tilde{C}_6 = 4$



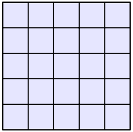
$t = 7$
$\tilde{C}_7 = 2$

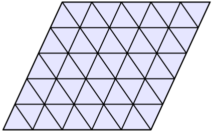


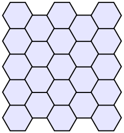
$t = 8$
$\tilde{C}_8 = 1$

Results - general perspective

1. **Result 1:** We prove that the maximum expected payoff for the selector who plays optimally with full information is, surprisingly, very close to the one for the optimal selector playing the blind game for any graph G with N vertices and maximum degree bounded from above by $o(\sqrt{N})$.
   (We are going to refer only to blind and full information games, as the expected payoff for the partial information game falls between those two.)
2. **Result 2:** We provide tight estimates for the maximum expected payoff in case when G is a square, a triangular or a hexagonal lattice.







Motivation and previous results

- The presented optimal stopping question may be treated as one of many generalizations of the celebrated **secretary problem** (consult [5] and [1]). One generalization consider graphs instead of partial orders as underlying structures.
- We continue the study when the underlying structure is a graph. Instead of maximizing the probability that the last vertex belongs to some previously defined set (which was the classical setup so far), we aim at maximizing the expected number of components at the moment of stop.
- The study of components is classical topic in the area of random graphs.

- The model we consider appeared for the first time in [4] by **M. Lasoń**. For G being a k-tree (k - constant) he proved:
    - there is no asymptotically better algorithm than wait until $\frac{1}{k+1}$ fraction of vertices;
    - the maximum expected number of components is then $\left(\frac{k^k}{(k+1)^{k+1}} + o(1)\right) N$;
    - asymptotically, the selector playing with full information does not get any advantage over the selector playing a blind game.
- The results from [4] are stated for k-trees, which are maximal k-degenerate graphs and maximal graphs with treewidth k. In contrast, 2-dimensional lattices are also k-degenerate but have unbounded treewidth. This motivates our particular study of lattices. (It turns out that the maximum expected payoff for 2-dimensional lattices is smaller than the one for k-trees in a non-negligible way.)
- The study of lattices is also motivated by the relation with the well researched site percolation problem on lattices.

Result 1 - detailed perspective

The following theorem shows how close are the optimal payoffs in full information and blind games when the maximum degree of G is bounded by $o(\sqrt{N})$. Recall:

$\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_N)$ - random permutation of vertices of G,
$\tilde{G}_t$ - graph induced by $\{\sigma_1, \sigma_2, \ldots, \sigma_t\}$,
$\tilde{C}_t$ - the number of components of $\tilde{G}_t$.

Theorem

Let G be a graph on N vertices. Let τ^f be the optimal algorithm while playing in a full information mode and let τ^b be the optimal algorithm while playing in a blind mode. For every $\varepsilon \in (0,1)$ there exists N_ε such that if $N \geqslant N_\varepsilon$ and the maximum degree of G is bounded by $D_{\varepsilon,N} = \frac{\varepsilon^2}{32}\sqrt{N}$, then

$$\mathbb{E}[\tilde{C}_{\tau^b}] \leqslant \mathbb{E}[\tilde{C}_{\tau^f}] \leqslant \mathbb{E}[\tilde{C}_{\tau^b}] + \varepsilon N.$$

The proof consists of three steps.

1. **Step 1**: considering different probabilistic model for the sake of more convenient computation.
   Let $p \in [0,1]$. Each vertex of G is declared *open* with probability p and *closed* with probability $1-p$, independently of the others. By G_p we denote the graph induced by the set of open vertices. C_p denotes the number of connected components of G_p. Letting $p = t/N$, where $t \in \{1,2,\ldots,N\}$ we expect that C_p and $\tilde{C}_t$ behave similarly when maximum degree of G is bounded by $o(\sqrt{N})$. In order to compare the random variables C_p and $\tilde{C}_t$ we use **coupling**.

Lemma

Let $t \in \{1,2,\ldots,N\}$. Let G be a graph on N vertices with the maximum degree bounded by D. Then

$$\mathbb{E}[C_{t/N}] \leqslant \mathbb{E}[\tilde{C}_t] + \frac{1}{2}D\sqrt{N}.$$

2. **Step 2**: proving that C_p is concentrated around its mean for certain G, in particular for G with maximum degree bounded by $o(\sqrt{N})$.

Lemma (C_p concentration)

Let $p \in [0,1]$. For every $\varepsilon \in (0,1)$ there exists N_ε such that if G is a graph with $N \geqslant N_\varepsilon$ vertices and $\sum_{j=1}^{N} \deg(v_j)^2 \leqslant \delta_\varepsilon N^2$, where $\delta_\varepsilon = \frac{\varepsilon^2/64}{\ln(64/\varepsilon^2)}$ then C_p satisfies

$$\mathbb{P}\big[C_p \geqslant \mathbb{E}[C_p] + (\varepsilon/8)N\big] \leqslant \varepsilon^2/64.$$

We prove this lemma using **Azuma's inequality** tailored for combinatorial applications (see [2] and [3]).

Lemma (Azuma's inequality)

Let $Z_1, Z_2, \ldots, Z_M$ be independent random variables, with Z_j taking values in a set Λ_j. Assume that a function $g : \Lambda_1 \times \Lambda_2 \times \ldots \times \Lambda_M \to \mathbb{R}$ satisfies, for some constants b_j, where $j \in \{1, 2, \ldots, M\}$, the following Lipschitz condition:

- *if two vectors $\mathrm{z}, \mathrm{z}' \in \Lambda_1 \times \Lambda_2 \times \ldots \times \Lambda_M$ differ only in j^{th} coordinate, then $|g(\mathrm{z}) - g(\mathrm{z}')| \leqslant b_j$.*

Then the random variable $X = g(Z_1, Z_2, \ldots, Z_M)$ satisfies, for any $t \geqslant 0$,

$$\mathbb{P}[X \geqslant \mathbb{E}[X] + t] \leqslant \exp\left\{ \frac{-2t^2}{\sum_{j=1}^{M} b_j^2} \right\},$$

$$\mathbb{P}[X \leqslant \mathbb{E}[X] - t] \leqslant \exp\left\{ \frac{-2t^2}{\sum_{j=1}^{M} b_j^2} \right\}.$$

To prove Lemma (C_p concentration) we apply Azuma's inequality in the following way. Let $p \in (0,1)$. For $i \in \{1, \dots, N\}$ and $V = \{v_1, v_2, \dots, v_N\}$

$$Z_i = \begin{cases} 1 & \text{if} \quad v_i \text{ belongs to } G_p, \\ 0 & \text{otherwise.} \end{cases}$$

Put $g(Z_1, Z_2, \dots, Z_N) = C_p$. For two vectors $\mathrm{z}, \mathrm{z}' \in \{0,1\}^N$ that differ only in j^{th} coordinate we have $|g(\mathrm{z}) - g(\mathrm{z}')| \leqslant \deg(v_j)$ unless v_j is isolated in G. For $j \in \{1, 2, \dots, N\}$ define

$$b_j = \begin{cases} \deg(v_j) & \text{if} \quad \deg(v_j) > 0, \\ 1 & \text{if} \quad \deg(v_j) = 0. \end{cases}$$

We get

$$\sum_{j=1}^{N} b_j^2 \leqslant N + \sum_{j=1}^{N} \deg(v_j)^2,$$

where N is the upper bound for the sum of squared ones while summing over vertices isolated in G.

3. **Step 3**: showing (using Step 1 and Step 2) that the values of $\tilde{C}_t$ are very likely to stay close to their expectation $\mathbb{E}[\tilde{C}_t]$ throughout the game.
   This implies that getting more information about $\tilde{G}_t$ does not help to achieve significantly better payoff than when playing blind.

Result 2 - detailed perspective

Summary of results obtained for 2-dimensional lattices.

Theorem

Let τ^b be an optimal algorithm while playing a blind game and τ^f be an optimal algorithm while playing a full information game on a lattice with N vertices. Let $c^b = (1/N)\mathbb{E}[\tilde{C}_{\tau^b}]$, $c^f = (1/N)\mathbb{E}[\tilde{C}_{\tau^f}]$. For sufficiently large N, we have:

Lattice	***Lower bound for c^b***	***Upper bound for c^b***
square	0.12953	0.13268
triangular	0.09629	0.10106
hexagonal	0.16738	0.17144

Furthermore, for every $\varepsilon \in (0,1)$ and for sufficiently large N, we have:

$$c^b \leqslant c^f \leqslant c^b + \varepsilon.$$

To obtain results for lattices we used two well-known facts about planar graphs:

1. if G is planar and connected, and $|E| > 1$ then 2|E|>3|F|,
2. Euler's formula for planar graph G: $|C| = |V| - |E| + |F| - 1$,

   where
   $|V|$ - number of vertices of G,
   $|E|$ - number of edges of G,
   $|C|$ - number of components of G,
   $|F|$ - number of faces of G.

Future work and open problems

1. Is it possible to relax the condition about the maximum degree in Result 1, e.g., to $o(N)$?
2. Closing the gap in the results for lattices (Result 2) would answer the open questions stated in percolation theory.
3. Considering presented problem for structures modeling social networks (e.g., preferential attachment graphs) would bring the research closer to the real-life applications.

References

Thomas S. Ferguson.
Who solved the secretary problem?
Statist. Sci., 4(3):282–289, 08 1989.

Wassily Hoeffding.
Probability inequalities for sums of bounded random variables.
J. Am. Stat. Assoc., 58(301):13–30, 1963.

Svante Janson and Andrzej Ruciński.
The infamous upper tail.
Random Struct. Algor., 20(3):317–342, 2002.

Michał Lasoń.
Optimal stopping for many connected components in a graph.
arXiv:2001.07870v1, 2020.

D.V. Lindley.
Dynamic programming and decision theory.
Appl. Stat. - J. Roy. St. C, 10(1):39–51, 1961.