The Forgetfulness of Balls and Bins

Will Perkins
Courant Institute

The Forgetfulness of Balls and Bins

Will Perkins
Courant Institute

ETH Mittagsseminar, 23/9/10

Thursday, September 23, 2010
Planted Substructures

Random Graphs:

Standard   Planted
Planted Substructures

Random Graphs:

Standard

Planted
Planted Substructures

Random Graphs:

Standard

Planted
Planted Substructures

Random Graphs:

Standard  Planted
Balls and Bins models:

Standard

Planted
Balls and Bins models:

Standard

Planted
Balls and Bins models:

Standard

Planted
Balls and Bins models:

Standard

Planted
Can we distinguish the Standard and Planted models?

- What does it mean to be able to distinguish two distributions?
- We use Total Variation distance
Can we distinguish the Standard and Planted models?

• What does it mean to be able to distinguish two distributions?

• We use Total Variation distance

\[ \text{TV}(P, Q) = \sum_{x: P(x) > Q(x)} P(x) - Q(x) = \sup_{A} P(A) - Q(A) \]
Total Variation Distance as a Game

• I choose distribution $P$ or $Q$ with probability $1/2$ each

• I draw a sample from the chosen distribution

• You guess which distribution it came from

• TV distance is your probability of success under the best possible strategy

\[ TV = 2p_s - 1 \]
Behavior depends on a parameter of the random model:

As you add more edges, it is easier to ‘hide’ the planting.
Behavior depends on a parameter of the random model:

As you add more edges, it is easier to ‘hide’ the planting
Optimal Strategy

By the Neyman-Pearson Lemma, optimal strategy given a sample $x$ is to compute $P(x)$ and $Q(x)$ and pick the larger

But we want something better:

- Simple, more descriptive statistic
- Asymptotic distribution
Example: For $G(n,p)$ with a planted triangle, the number of triangles is the distinguishing statistic and is asymptotically Poisson

$$p = \frac{c}{n}$$

Standard model has $\frac{c^3}{6}$ expected triangles. Planted model has $\frac{c^3}{6} + 1$
Example: For $G(n,p)$ with a planted triangle, the number of triangles is the distinguishing statistic and is asymptotically Poisson

$$p = \frac{c}{n}$$

Standard model has $\frac{c^3}{6}$ expected triangles. Planted model has $\frac{c^3}{6} + 1$

Best strategy: count triangles, and pick a distribution with closest mean.
Example: For $G(n,p)$ with a planted triangle, the number of triangles is the distinguishing statistic and is asymptotically Poisson

$$p = \frac{c}{n}$$

Standard model has $\frac{c^3}{6}$ expected triangles. Planted model has $\frac{c^3}{6} + 1$

Best strategy: count triangles, and pick distribution with closest mean

TV distance in critical window is

$$\text{TV (Pois}(\mu), \text{Pois}(\mu + 1))$$
Balls and Bins:

• Plant \( n \) balls in \( n \) bins in a fixed configuration, throw \( m-n \) on top at random
• \( a_i \) is the number of balls planted in bin \( i \)
• \( z_i \) is the number of balls that end up in bin \( i \)

\[
a_2 = 1, \quad z_2 = 3
\]
Questions:

• Given an initial planting, what is the critical scaling \( m = m(n) \) so that the TV distance between the Standard and Planted distributions is non-trivial?

• At the critical scaling, what is the TV limit?

• What is the optimal strategy for the TV game?

• What is the distinguishing statistic?
Two Extreme Examples:

- $n$ balls planted in one bin
Two Extreme Examples:

- $n$ balls planted in one bin

Distinguishing statistic?

Number of balls in bin #1
Two Extreme Examples:

• $n$ balls planted in one bin

Distinguishing statistic?

Number of balls in bin #1

• one ball planted in each of $n$ bins
Two Extreme Examples:

- $n$ balls planted in one bin

Distinguishing statistic?
Number of balls in bin #1

- one ball planted in each of $n$ bins

Distinguishing statistic?
Two Extreme Examples:

- $n$ balls planted in one bin

Distinguishing statistic?
Number of balls in bin #1

- one ball planted in each of $n$ bins

Distinguishing statistic?
One ball in each bin? No. Number of pairs.
Method

• Write out exact formula for ratio of probabilities

\[
\frac{PL(Z)}{ST(Z)} = \frac{n^n(z_1)^{a_1} \cdots (z_n)^{a_n}}{(m)_n}
\]

Under ST distribution, this is a RV with mean 1

• Find concentration Threshold

• Compute Asymptotics
Results:

Two regimes: ‘Flat’ and ‘Hilly’
Results:

Two regimes: ‘Flat’ and ‘Hilly’

• Flat regime: Statistic is number of pairs of balls in the same bin. Critical scaling is

\[ m \sim cn^{3/2} \]
Results:

Two regimes: ‘Flat’ and ‘Hilly’

• Flat regime: Statistic is number of pairs of balls in the same bin. Critical scaling is

\[ m \sim cn^{3/2} \]

• Hilly regime: Statistic is number of balls in each bin weighted by the number of planted balls:

\[ \sum_{i=1}^{n} a_i z_i \]
Results:

Two regimes: ‘Flat’ and ‘Hilly’

• Flat regime: Statistic is number of pairs of balls in the same bin. Critical scaling is
  \[ m \sim cn^{3/2} \]

• Hilly regime: Statistic is number of balls in each bin weighted by the number of planted balls:
  \[ \sum_{i=1}^{n} a_i z_i \]

Critical scaling is a function of ‘hilliness’. Ranges from \[ m \sim cn^{3/2} \] (nearly flat) to \[ m \sim cn^3 \] (all \( n \) in one bin)
Further Questions:

Unlabeled bins: To find critical scaling need to understand concentration of sums of log-normal random variables

$$\sum \exp \left( \sum \sigma a_{\sigma(i)} N_i \right)$$
Further Questions:

Unlabeled bins: To find critical scaling need to understand concentration of sums of log-normal random variables

\[ \sum \exp \left( \sum_{i} a_{\sigma(i)} N_i \right) \]

Similar sums appear in study of Spin Glasses
Further Questions:

Unlabeled bins: To find critical scaling need to understand concentration of sums of log-normal random variables

$$\sum_{\sigma} \exp \left( \sum_i a_{\sigma(i)} N_i \right)$$

Similar sums appear in study of Spin Glasses

What are exact distinguishing thresholds of large planted subgraphs in $G(n,p)$?
Further Questions:

Unlabeled bins: To find critical scaling need to understand concentration of sums of log-normal random variables

\[ \sum_{\sigma} \exp \left( \sum_{i} a_{\sigma(i)} N_i \right) \]

Similar sums appear in study of Spin Glasses

What are exact distinguishing thresholds of large planted subgraphs in \( G(n,p) \)?

On what other random models could we use the same type of analysis?
Thank You!