Instructions (updated on 23 August 2012)

A. Overview of the Challenge.

Pay-per-click is a popular payment model for Internet advertising where the *advertisers* contract a *syndicator* to distribute their advertisements to the *publishers.* The syndicator keeps track of the number of user clicks on the advertisement displayed in the publisher's site, and the publishers are paid according to the number of clicks through their sites.

Pay-per-click is subjected to abuses by malicious publishers through *click fraud* where the publishers imitate legitimate users, or mislead the users to generate clicks that do not have genuine interest in the advertisements' content. By maliciously inflating the number of clicks, the advertisers are wrongly overcharged and such fraud is considered a crime in many jurisdictions.

The following is a typical sequence of clicks traffic received at a syndicator:

```
|time           | user's ip    |publisher's id|advertisement's id|
|02/01/12 12:01|123.123.123.12|   100        | 129              |
|02/01/12 12:01|123.123.123.12|   100        | 423              |
|02/01/12 12:01| 93. 93.100.22|   530        |  35              |
|02/01/12 12:01|123.123.123.12|   100        | 533              |
|02/01/12 12:01|123.123.123.12|   100        | 153              |
|02/01/12 12:02| 89.109. 23.50|   219        | 357              |
|02/01/12 12:02|123.123.123.12|   100        | 423              |
```

The clicks from the publisher id 100 are suspicious as it is unlikely that a legitimate user generates many clicks within a second. In practice, there are more sophisticated patterns, and there are more information in the click traffic (for instances, the user's browser type and the "referer"), and more information on the publishers and advertiser (for instances, the types of advertisements and the publishers' bank accounts).

By entering into this contest, you are taking up the challenge in solving a real-life problem: deriving an automated method that identifying malicious publisher from large volume of clicks. The data (with certain fields anonymized for privacy protection) are provided by BuzzCity, a global advertising network.

After successfully registered, you will have access to the *Training Data,* which consists of the clicks traffic on selected days over two weeks, and the identity of the malicious publishers. Your program will be tested on the *Test Data*, which are clicks received a week after the Training Data are gathered. Your program may uses parameters extracted from the Training Data, which can be stored as a file to be read by the program. On input of the Test Data, your program has to output a list of malicious publishers. The performance of your program is measured by a quantitative measure based on the precision and recalls of the output.

B. Eligibility

Open to all current SOC students as at 1 August 2012.

C. Registration Instructions.

Registration information is available at the Challenge webpage:
http://www.comp.nus.edu.sg/~clickfrd/
At most 3 members per team. A contestant can be registered in one team only.

Every member in the team has to provide his/her SOC Unix account name. A confirmation email will be sent to the Unix account of each registered contestant.

Each registered contestant can access the Challenge online-forum site, which is hosted in IVLE under the module code OTH150 and module name "BuzzCity ClickFraud Detection Competition". Access to the module will be granted to a contestant on the next working day after the contestant has successfully registered.

The Training Data can be downloaded from the workbin in IVLE.

D. Platform.

Your program will be tested on Sunfire.

E. Submission Instructions.

The submitted (executable) program must be stored in the directory
`/homedirectory/click/bin`
where *homedirectory* is the team leader's home directory in the SOC Unix account. The filename of the submitted program must be
`/homedirectory/click/bin/detect`

The source files of the submitted program must be stored in
`/homedirectory/click/src`
The directory must contain a file `README` describing how the executable program
`/homedirectory/click/bin/detect`
is obtained from the source files. The judges may ask the team to demonstrate the compilation process during the judging period. A team will be disqualified if it is unable to demonstrate the compilation process during the judging period.

The program may read/write file in the directory
`/homedirectory/click/bin`

(Input format) The input format will be announced in IVLE.

(Output format) The program must write to the standard output a sequence of publishers' identities. Example of the output format will be announced in IVLE.

F. Judging Criteria

1. The performance of your program is calculated based on the F1-score (also known as F-measure) of the publishers identified as malicious by your program. The F1-score consider both the precision and recall. Let
   X be the set of malicious publishers in the Test Data, and
   Y be the set of publishers identified by your program,
   then
   $precision = |Y \cap X| / |Y|$,
   $recall = |Y \cap X| / |X|$,
   and the F1-score is defined as
   $2 (precision \cdot recall) / (precision + recall)$.

   During the Question and Feedback phase, the criteria may be modified.

2. Your program must not consume unreasonably large computing resources. The time limit and memory requirements will be determined later and updated in this document.

3. Essentially, the judges will copy the whole directory
   `/homedirectory/click/bin/`
   to a working directory and then execute the file `detect` in the working directory.

G. Timeline and Important Dates

4. 29 Aug 2012. **Briefing**. See the Challenge page for venue and changes of date/time.

5. 15 Aug 2012 to 15 Oct 2012. **Registration**.

6. 15 Aug 2012 to 15 Oct 2012: **Question and feedback phase**. During this period contestants may post questions and requests in the IVLE forum page. The organizer may modify the judging criteria, update or provide more data during this period. After 15 Oct, all criteria will be fixed and the organizer will not provide further information of the data.

7. 2 Jan 2013 to 7 Jan 2013: **Verification phase**. The organizer will conduct a test run on 2 Jan 2013. That is, the organizer will test the submitted on the sample test data. Contestants may still modify their programs.

8. 7 Jan 2013 to 11 Jan 2013: **Judging phase**.
   During this period, the judge may contact the contestants for clarification. If the contestants are not contactable during this period, the judges will make decision based on whatever being submitted.

9. 12 Jan 2013: **Announcement of winners**.

10. 19 Jan 2012: **Presentation and Award Ceremony**.

H. FAQ

1. *Can I participate in more than one team?*
   No.

2. *Can I join a team after the team has registered?*
   Yes.

3. *I had withdrawn from a team. Can I now join another team?*
   You need permission from the judges. The judge may seek the opinion of the team you had withdrawn from.

4. *Can I share the data with friends who are not taking part in the contest?*
   See the Terms-and-Conditions. Note that the data is licensed under the Creative Commons Attribution 3.0 Unported License.
   Please refer to:
   http://docs.buzzcity.net/wiki/BuzzCity_Advertising_Dataset_-_Creative_Commons_License