Information extraction from Facebook profiles

Summary

In the html source of a Facebook page there is a lot of unknown data. Our goal was to research a specific list of Facebook peoples’ IDs in the source html. The list appears in the same place in the html page, but the order of the IDs may change from time to time. Our first finding is that the list consists of friends only. The second finding and the most important one, is about the order of the IDs in the list. We discovered that people that are “related” to each other appear higher on each other’s list. Therefore, interactions in Facebook such as Like, commenting, chatting and more, can move an ID to a higher place in the list. For example - for three people A, B and C, if A appears higher on C’s list than B, then A is “closer” (according to Facebook) to C than B. we also introduce a new way of describing Facebook friends’ networks. Not as an undirected graph but as a directed weighted graph. Where there is an edge from person A to person B, as if Bs’ ID appears in As’ list. The weight is the index of B’s ID in A’s list.

To prove our hypothesis (the correlation between the relations in Facebook and the order of the list), we used a small network of Facebook users which we created for this task. We made interactions between a specific user (the test user) and all the others and recorded the changes in the test user’s list.

We also created a software to automatically create the directed weighted graph from a given html pages where in the graph there are names instead of IDs. The extractions of names from the ids is slow and Facebook can detect the software as a bot. To avoid these problems, we used AWS to distribute the computation and make it scalable. So, if a specific computer extracts less names, its probability to succeed in avoiding detections is growing.