QUIRK'S
Marketing Research Review

Why do apps want my sensitive data?

| By Michael Lieberman

Data privacy has been at the forefront of the legislative battle waged by marketing research organizations like the Marketing Research Association and ESOMAR. This issue looms large in the media as well. Advances in technology only make the matter more urgent.

Privacy concerns are linked to the exploding number of Android and Apple apps on smartphones and tablets. When one downloads a newspaper, game, etc., that application requests permissions by default. Most of us barely glance at what we agree to. Are these permissions necessary for the application to run? Are they an invasion of privacy? There is a hidden tax in our mobile apps and that debt is paid in the coin of privacy. It is hard to differentiate among the legitimate, superfluous or even malicious permissions.

Google requests permission for more than 200,000 apps. In this article I will examine and present results of a Gmail segment study and suggest a model for identifying anomalies in permission requests made by applications.

When one downloads an application, Android usually suggests a similar app, much like Amazon might suggest a book after you make a purchase. The suggestions are derived by a formula. This is similar to a priori market segmentation. In other words, certain apps go with other apps.

Google makes its categories and data public. Google shares whether each app requests one or more of up to 189 permissions. Most of them are quite common. My team downloaded close to 189,000 data points (apps) in our database. These are provided in comma-separated value format, which is easily read in Excel.

Permissions hypothesis

It is reasonable to assume that Google segments apps for a reason. After all, suggested apps – be it a calorie counter, a simulator war game or a memory exerciser – are designed to carry out certain functions. Thus, one would expect that the permissions granted to similar apps would be similar. It is against this hypothesis that we can search for unusual permissions.

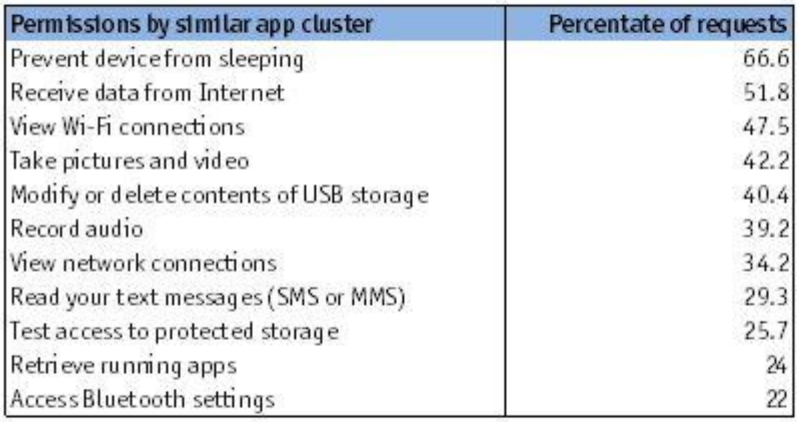
Computer programmers have approached me to find a one-size-fits-all algorithm to weed out bad permissions. But bad is a relative term. For example, it would be bad if I had the codes to the U.S. nuclear arsenal but not necessarily bad if the president had access to those same codes. It is good if I know the password for my bank account, bad if someone else does.

My hypothesis is that the apps contained in the Gmail cluster will have similar permissions. I will be looking at exceptions to that rule to decide if any could be labeled as a bad permission.

Similar apps

Table 1 lists some of the aggregated permissions that are requested within the Gmail cluster. The most frequent permissions are shown in descending order by the percentage of requests within each similar application group.

Table 1



Permissions by similar app cluster	Percentate of requests
Prevent device from sleeping	66.6
Receive data from Internet	51.8
View Wi-Fi connections	47.5
Take pictures and video	42.2
Modify or delete contents of USB storage	40.4
Record audio	39.2
View network connections	34.2
Read your text messages (SMS or MMS)	29.3
Test access to protected storage	25.7
Retrieve running apps	24
Access Bluetooth settings	22

Naturally, some permissions float to the top. This study is seeking the uncommon ones. Glancing at a simple output is not enough. Further analysis is needed.

Network analysis

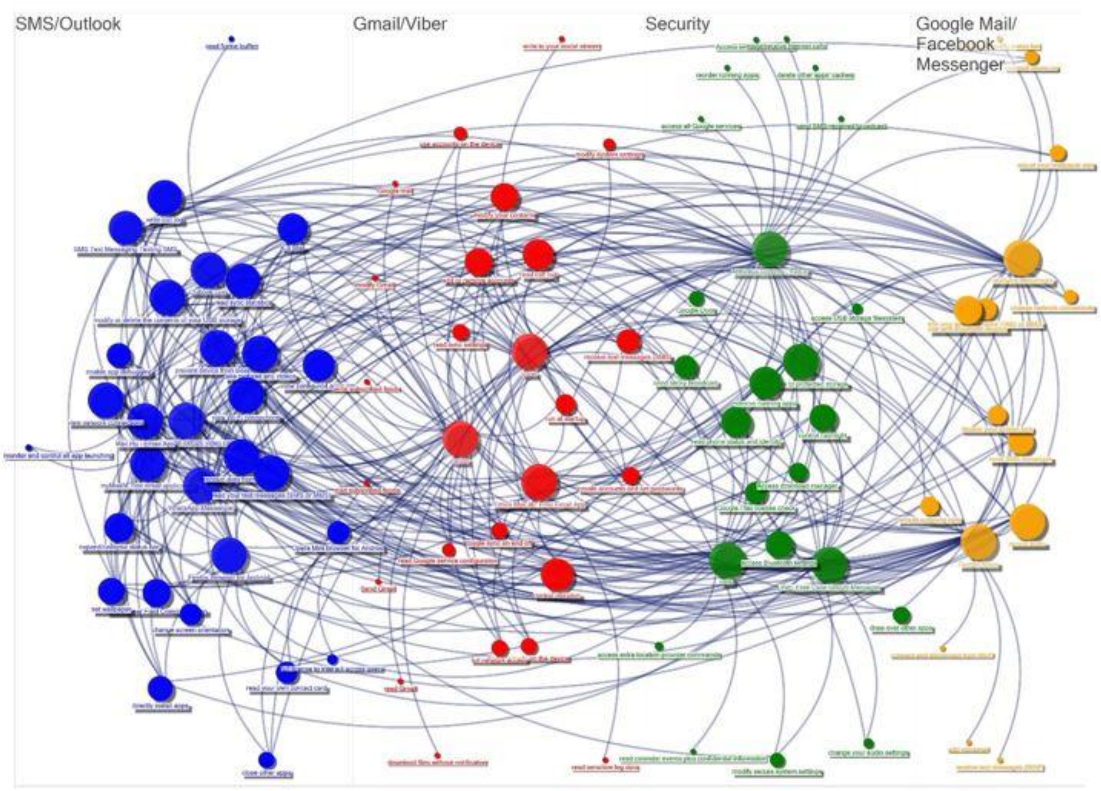
Network analysis is the practical use of graph theory. Graphs can be used to model many types of relations and processes in physical, biological, social and information systems. Many practical problems can be represented by graphs.

In computer science, graphs are used to represent networks of communication, data organization, computational devices, flow of computation and so on. For instance, the link structure of a Web site can be represented by a directed graph, in which the vertices represent Web pages and directed edges represent links from one page to another. A similar approach can be taken to problems in travel, biology, computer chip design and many other fields.

There are many statistical measures calculated in a network analysis. For our practical purposes, only two measures need to be subjected to our analysis: degrees, how popular a given node is, and eigenvector, how well connected a node is to other clusters. This is called a bridge.

The first step is to structure the data on which apps are connected to which permissions. This can be done using NodeXL, an open-source Excel back-end program. Together with IBM SPSS, the data is shaped, placed into Excel and run with the analysis, producing the following graph (Figure 1).

Figure 1: Network Analysis of Gmail-Similar Apps Cluster and Permissions Requested



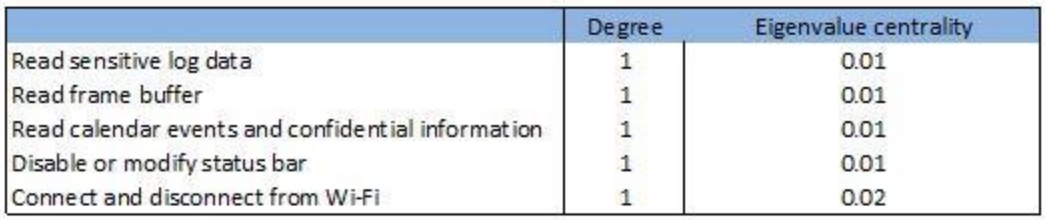
The size of each app and permission is the eigenvector centrality – its bridge factor. Despite being classified as similar applications, the analysis has isolated four distinct clusters within the group. These clusters, shown in Figure 1, are not segmented by Google but rather by which permissions the apps share. Not unexpectedly, the clusters divide into sub-segments of Gmail and Gmail-like apps.

Assumptions reveal exceptions

Network analysis is useful when examining all kinds of data, such as Twitter and Web site links, restaurant data and media research. The general approach is to examine the top degree (popularity) and eigenvector centrality (bridge). These key measures suggest a sales strategy.

Here, though, we are looking for exceptions. To do so, we reverse the degree and eigenvector centrality, searching for the smallest values within the Google group. When we reverse the network analysis statistics, we see that some permissions have low degree and eigenvector centrality, which means they are uncommon (Table 2).

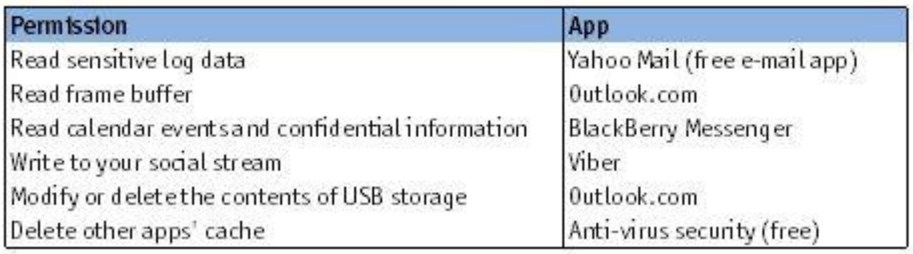
Table 2



	Degree	Eigenvalue centrality
Read sensitive log data	1	0.01
Read frame buffer	1	0.01
Read calendar events and confidential information	1	0.01
Disable or modify status bar	1	0.01
Connect and disconnect from Wi-Fi	1	0.02

Table 3 provides a smattering of information about the isolated permission findings from the network analysis.

Table 3



Permission	App
Read sensitive log data	Yahoo Mail (free e-mail app)
Read frame buffer	Outlook.com
Read calendar events and confidential information	BlackBerry Messenger
Write to your social stream	Viber
Modify or delete the contents of USB storage	Outlook.com
Delete other apps' cache	Anti-virus security (free)

Glancing at Table 3, one might ask the following questions:

- For those of us who still use Yahoo Mail, do we want it to read our sensitive log data?
- Are we comfortable with Outlook taking a screenshot of other apps programmatically, without root permission (frame buffer)?
- BlackBerry looks to be on its way out but while you were still using the device it may have read confidential information. What was it doing with the data?
- Anti-virus security might be a little too secure. We give it permission to automatically delete other apps' cache. Which apps? What cache?

How should we, as market researchers, respond to app permissions and overall data privacy? What should our role be, as data collectors? Generally, I wrap up my articles with useful tips. Today I leave you with more questions than answers. This reflects the reality of our information age.

Editor's note: Michael Lieberman is founder and president of Multivariate Solutions, a New York-based statistical consulting firm