A view of the world through the eyes of Wikipedia

Introduction
This paper analyzes the relationships between world countries by examining a graph of co-occurrences in the articles of Wikipedia, in the attempt to answer the following question: what are the most powerful factors that bring two country names together in encyclopedia articles – is it geographical proximity, economy, political regime, spoken language, or common history?

Data source
The data was gathered from the Wikipedia main site, which provides a download location for the entire encyclopedia article repository: http://en.wikipedia.org/wiki/Wikipedia:Database_download. In order to define the graph, approximately 5000 of the paragraphs from the database containing at least two countries were sampled was used to define edges.

Because of the multiple meanings of some country names (like Turkey or Georgia), those were filtered out from the graph.

Headlines
USA: Axis Mundi
One first insight we discovered was that the country mentioned in most articles in Wikipedia is USA. Not only that, but it’s also the country mentioned most in relation with others. See Figure 1 below.

Figure 1: USA: The center of the world. In the graph, each vertex represents one country. An edge was drawn between two countries if out of the sampled paragraphs, there were at least 4 containing both of them. Node sizes are proportional to the number of occurrences of the country name in the sampled paragraphs, and edge thickness is proportional to the number of co-occurrences of the two countries together:
Wikipedia editors leave out the under-developed part of the world

We further applied forces and re-organized the graph, coloring by geographical regions of the world. We discovered a direct relationship between how developed a country is and the number of its occurrences in Wikipedia articles. Thus, the countries that are most frequently mentioned are: United States, Canada, Germany, France and United Kingdom. There are also very strong edges between them. The United Kingdom, Germany and France form an attractor pole for all other European countries, which seldom are mentioned alongside countries from other parts of the world. The same trend, though not so evident, can be noticed in the case of Asian countries, with China and Japan.

The under-developed countries from Africa, Central American, and South America cluster together naturally based on their geographic proximity, having weak connections to the rest of the world, or in some cases, even with each other.

The obvious explanation for this is that Wikipedia contributors are, in their vast majority, from developed countries, and only write articles about their own countries, with little to no interest paid to the rest of the world.
Figure 2: Countries clustered according to Wikipedia co-occurrences. Node sizes and edge widths are computed like in Figure 1. Colors delimit clustering. Notice how all African countries cluster together, separated from the rest of the world. Our assumption is that the reason they are not entirely cut away from the developed countries in the encyclopedia is because the recent conflicts in Syria, Egypt and Libya.
Military conflicts: one of the best sources of correlation between countries in Wikipedia

One final thing we noticed was that the strongest connections between developed countries and clusters of third world countries are accomplished through only a few nodes, those representing places where there are or have been conflicts in the recent past: Libya, Lebanon, Egypt, Syria, and others. Also, the edges between the USA and Afghanistan, Iraq, and Iran are stronger than those between each of them and the rest of the Arab world.

We thus draw the conclusion that war brings countries in the same encyclopedia article more than geography or culture.

Critique of NodeXL

I found NodeXL to be a good tool for extracting insights. It offers good ways of examining networks of relationships. However, it still has some issues that make it hard for a user to make the best of its capacities:

- Twitter and Facebook importing: for Twitter, one needs a privileged account in order to be able to extract enough nodes to observe insights; for Facebook, the existing plugin is still unstable – in my case, it crashed every time I tried to retrieve data slightly larger than 100 nodes.
- The “Lay Out Again” functionality: it would be useful if there was a Stop button somewhere, and a progress bar, for the cases when the lay out algorithm takes too much time. My computer blocked a few times because of that, and the only way to solve the problem was by restarting Microsoft Excel.
- By default, the lay out algorithms apply to all of the vertices and edges in the spreadsheet, and in order to run them for just the filtered data, one needs to select that option every time from a drop-down box. In my opinion, the latter should be the default functionality.