The vocal minority: Local self-representation and co-editing on Wikipedia in the Middle East and North Africa

Bernie Hogan, Mark Graham,
Ahmed Medhat Mohamed
Oxford Internet Institute
University of Oxford
1 St. Giles, Oxford UK
Bernie.hogan@oii.ox.ac.uk

ABSTRACT
To what extent are people representing places to which they have some biographical information? Whose voice is influential? We analyze the geotagged articles in the English language Wikipedia. After filtering to the articles that georeference places in the Middle East and North Africa, we construct a co-editing network of authors. We then examine the talk pages of these authors to assess the self-declared locational affiliations of the authors (i.e. where they live, work or were born). We demonstrate that there exists few authors claiming to be from the MENA region, except for Israel, Iran and to a much lesser extent Egypt. This raises significant concerns about how Wikipedia represents the MENA region, if there are more self-declared editors from the UK or Canada than all MENA countries combined.

Categories and Subject Descriptors
J.4 SOCIAL AND BEHAVIORAL SCIENCES,

Keywords
Social Media, Wikipedia, Social Networks, Semantic Networks, Network Analysis, Visualization,

1. INTRODUCTION
Wikipedia fashions itself as a collaborative global reference site. Although Wikipedia is global in scope, it is not necessarily global in voice. We examine the network structure of individuals who self-describe as being from and write about the Middle East and North Africa (MENA region) on Wikipedia. We find that in virtually all of the 20 countries of interest, there is a single connected component of editors, alongside a small number of isolates. The size and shape of this component varies dramatically, however. In some countries, the component is very diffuse, indicating that no single individual or cluster is linking the network together. In other countries, the network is very densely connected, indicating a focus on a single article or small set of articles that define a key point of entry for people from that country.

The key differences in the English language version point to the stark differences in the extent to which individuals believe it is important to stake a claim on the English platform. While Israel and Iran appear to have a deliberate and concerted strategy of self-representation that shows through Wikipedia, most of the Arabic speaking countries do not. Consequently, there is very little local involvement from self-described local actors. Such lack of involvement may serve to reinforce stereotypes, myths or a perception of that area that is in contrast to the self-perception of those from the area.

To perform this analysis, we analyzed the entire Wikipedia corpus in English from October 2011, first finding articles that were geotagged and then falling within the MENA region. We then examined the history of all these articles building a two-mode network of authors that co-edit an article. We examined the talk pages of these authors and used NLP techniques to determine through contextual grammars whether an individual could be considered from, born or having worked in a particular country. This analysis demonstrates the potential for network analysis on unstructured data such as that found in Wikipedia free text.

2. Overview

2.1 Geographic differences
Not every article on Wikipedia is geotagged, nor could every article be a candidate for geotagging. However, across most language versions, nearly a quarter of all articles have some form of latitude and longitude embedded in the article, highlighting the importance and salience of spatial knowledge within this reference work. Past work has indicated that Wikipedia is commonly found at the top of Google search queries [1]. Generalized claims about Wikipedia being 'the' top search hit are now rendered problematic by personalized search results. Nevertheless, as one of the world's most popular sites, and unambiguously the world's most popular reference site, it is safe to assume Wikipedia is presently still a preeminent source for information about place.

In this work, we have discovered 708,002 geotagged articles in English in the entire Wikipedia corpus (of just over 4 million articles). Of these 21,297 articles were places within our area of interest. For these articles, 76394 authors have contributed under a user name, and 12,248 of them have been ascribed at least one country. To note, individuals could be ascribed to multiple countries as they may be born in one, work in another and report having lived in others still. Of those 12k authors, 2267 of them report an association with the UK, 1160 with Australia, 1106 with Mexico, 1049 with Canada. The largest number from the MENA region is Israel, with 421 authors. To reiterate, these are not authors in the total Wikipedia, but only those who have worked on place-oriented articles in the MENA region.

To create co-editing networks, we employ a novel algorithm that takes the varying length of articles into account.

Algorithm 1 Edge Weighting Algorithm for Unimodal Projection

input: Weighted Bipartite Contributor-Article Network B
output: Projected Contributor Network P

$C \leftarrow Contributor\ Nodes\ in\ B$
for c in C **do**
  $E \leftarrow edges(c)$
  for w,article in E **do**
    $E^* \leftarrow edges(article)$
    $N \leftarrow length(E^*)$
    for $w^*, c^* in\ E^*$ **do**
      if $c < c^*$ **then**
        if $P_{c,c^*}\ not\ in\ P$ **then**
          $P_{c,c^*} \leftarrow 0$
        end if
        $P_{c,c^*} \leftarrow P_{c,c^*} + \frac{w+w^*}{N-1}$
      end if
    end for
  end for
end for
return P

We then capture the subset of authors for each of the countries in our area of interest. Five example networks are shown in Figure 1.

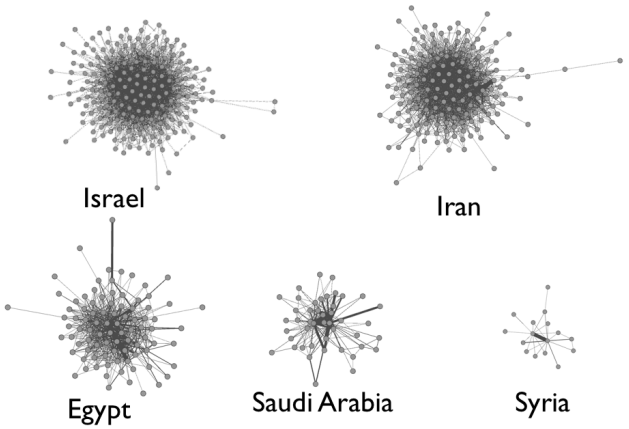

Figure 1. Co-editing network of authors who have self-declared being from a specific country, five examples.

To qualitatively summarize the resulting 20 co-editing networks, we can note that virtually all show a core periphery structure, with few isolates and all are dwarfed by the size of the co-editing networks from individuals ascribed to the west.

2.2 Caveats

We have a series of significant caveats to this work at this time. Most notably, our location gazeteer is very incomplete for the USA. Thus, we have only picked up on people to who explicitly claim to be from America, where most Americans will refer to either their city or state instead. We do not have this issue outside the US, as the gazetter used was, ironically, from the US and designed for foreign cities, states and countries.

Country	Nodes	Edges	Trans.	Giant C.
Israel	209	2412	0.42	198
Iran	198	1900	0.43	165
Egypt	112	678	0.47	89
Morroco	107	115	0.26	56
Lebanon	69	378	0.48	60
Saudi Arabia	54	187	0.41	45
Iraq	52	103	0.47	35
Tunisia	50	44	0.24	27
Sudan	39	9	0.13	8
Emirates	36	22	0.26	17
Syria	25	25	0.22	16
Palestine	21	68	0.55	21
DZA	20	27	0.71	10
Jordan	17	10	0.58	7
Kuwait	15	35	0.55	14
Bahrain	14	32	0.79	12
Libya	14	27	0.53	13
Yemen	8	11	0.44	8
Oman	7	3	0.00	4
Qatar	6	8	0.53	6
Comoros	2	0	0.00	1

Table 1. Network structure statistics for self-declared editors from countries in our area of interest (MENA region), displaying nodes, edges, transitivity statistics and size of giant component.

A second caveat is that editors from the middle east may be uncomfortable declaring as such. We are not asserting this is the full network, but the 'exposed' self-declared network. Finally, we are working on a parallel analysis with a gazeteer for Arabic.

Given the interest of geographic actors in 'self-focus' editing patterns (where one can perceive homophily by geographic location) [2] we have serious concerns about the extent to which geographic articles in the MENA region truly represent the ideas and issues of those in the region. We are exploring ways to highlight and facilitate content in this region.

3. REFERENCES

[1] Graham, M. Wiki Space: Palimpsests and the Politics of Exclusion. Lovink, G & Tkacz, N. eds., *Critical Point of View: A Wikipedia Reader*. 269-282. Institute of Network Cultures, Amsterdam, 2011.

[2] Hecht, B., Gergle, D. On the "Localness" of User-Generated Content. In *Proc. CSCW 2010*, ACM Press, (2010).