Quantitative intercultural comparison by means of parallel pageranking of diverse national wikipedias

Daniel Hromada

Ecole Pratique des Hautes Etudes / CHART / Lutin Userlab

Abstract

The aim of our study was to show that distributions of hyperlinks within wikipedia corpora implicitly contain information about cultural preferences of its authors. We have transformed wikipedia corpora written in 27 different languages into graph structures whose vertices correspond to wikipedia articles and edges to hyperlinks between these articles. Afterwards we have calculated PageRank vectors for every one of these graphs, thus obtaining so-called "intracultural importance list" for every linguistic community under study. Two datamining experiments were performed with obtained data: "the top country" study indicated that labels of articles concerning countries, related to linguistic community that created these articles are to be found in the top parts of their respective intracultural lists and inversely that the top parts of these lists can be potentially used as a stylometric method of identification of the community which created the corpus. "The world&corpus" study revealed that majority of rankings of articles concerning the countries of reference within intracultural list of a given community significantly correlates with a factual geographic distance between the country of reference and a supposed home country of a linguistic community. Both experiments have indicated presence of morphism between wikipedia hyperlink graph and a factual world of its authors.

Keywords: PageRank, Wikipedia, graph theory, comparative culturology, quantitative anthropology, cultural stylometry, world-corpus correlations

1. Introduction

The aim of this article is to propose a new quantitative method for comparison of different cultures by reducing culture-specific corpora to a common metrics. We shall try to demonstrate the feasibility of such an approach by using PageRank as such a metric and wikipedias of diverse (mostly European) linguistic communities as corpora which will be compared.

Both Wikipedia and Pagerank have lately received a substantial amount of attention from different scientific fields. Considered by some to be «probably the most important single contribution to the fields of information retrieval and Web search of the last ten years » (Esuli and Sebastiani, 2007) implementation of PageRank by (Brin and Page, 1998) was without a doubt a key component of ascent of Google to the very top of most visited Internet sites.

On the other hand, Wikipedia is based upon a very simple idea of self-organized collaboration of a huge number of authors. The hypothesis that such a huge number will, in the long run, approximate scientific truth better than a limited number of experts (Surowiecki, 2004) is far from being ultimately proven. However, Wikipedia is nowadays considered as reliable source of information in many domains, and it is one of the most important and freely available encyclopaedic corpora. Its multilingual properties are being more and more exploited in NLP

research for sense disambiguation word sense disambiguation (Mihalcea, 2007), question answering (Ferrandez et al., 2007), named entity recognition (Richman and Schone, 2008).

Only few studies, however, focused fully upon differences between diverse wiki corpora. And even when such "exploiting asymmetries" (Filatova, 2009) or "information arbitrage" (Adar et al., 2009) were presented, their goal was to infer data from article-content related discrepancies, and not to make comparisons between corpora considered as consistent wholes.

Research presented by this paper aims to demonstrate that even such large-scale comparisons can yield valid information. Our starting hypothesis can be stated like this: *Wikipedia maybe does not approximate scientific truth, but it certainly approximates culture of its authors*. In more exact terms, supposing that 1) the very act of creation of an article or a link presupposes an existence of a biased preference within the author and 2) that wikipedia is a graph structure whose vertices are equivalent to articles and edges to hypertext links between this articles, we propose that such a graph is at least partially but significantly isomorphic with associative network of culturally determined meanings and values of its authors.

Proposal that culture – which can be conceived as structure of symbols, artifacts, buildings, institutions, social roles etc. which are mutually interconnected in a very specific way– can be described by graph theory and later analyzed by network analysis is far from being new (for an overview, see Park, 2005). Validating such a hypothesis, however, is not easy since it is not easy to find a 1) unique graph-like structure (*e.g*. structure with vertices and edges) that 2) represents common activity of huge number of culture-holders. And even when such a structure is found, the question whether it faithfully represents (is isomorphic with) a given culture is difficult to answer.

But since it is nowadays widely accepted that culture is in the first place distinct from other cultures and that this distinction forms the very essence of a given culture (Bourdieu, 1979), even when it is almost impossible to compare a cultural graph with factual world itself, cultural graphs can always be compared with each other and the results of this comparison can be subsequently more easily compared with evident cultural distinctions of factual world.

We propose that corpora of local wikipedias created by diverse linguistic communities can serve as a basis for construction of such «cultural graphs» and that these graphs can be subsequently compared by means of PageRank centrality measure.

2. "The top country" study

Since a "corpus culturology" doesn't seem to be an explored scientific domain, the goal of this preliminary analysis was to decide whether it is worth to continue with implementation of more robust statistic techniques or whether to consider as false the very introductory hypothesis "hyperlink distribution of a wikipedia graph contains implicit information about cultural preferences of its authors". In other words, our primary intention was to assess whether some culture-specific information can be observed by applying a PageRank algorithm on wikipedia corpora of diverse linguistic communities.

2.1. Method

Database tables «pages» (containing the list of articles – vertices) and «pagelinks» (containing the list of hypertext links – edges) were downloaded from wikimedia's site.

All vertices and edges not having namespaces 0 (article) 14 (category) and 100 (portal) were removed from the tables; subsequently a *page_from* → *page_to* plaintext edge list was

generated. After this edge list was transformed into a graph G, pagerank vector – which is in fact the eigenvector of graph's modified adjacency matrix – was calculated by igraph library (Csárdi and Nepusz 2006). Damping factor d=0.77 was chosen for the calculation. These transformations and calculations were repeated for 27 wikipedia corpora, overall properties of their respective graphs are present in Tab. 1.

ISO 639 code	*Name of language*	*Number of vertices (articles)*	*Number of edges (hyperlinks)*
AR	Arabic	234538	4963998
BG	Bulgarian	143439	3578973
CS	Czech	266854	7187995
DA	Danish	205245	4402963
DE	German	1939647	43782766
EL	Greek	82168	1879300
ES	Spanish	1303273	23212253
ET	Estonian	126448	2580511
FI	Finnish	403380	7609470
FR	French	1996383	53003962
HE	Hebrew	245431	9103883
HR	Croatian	116515	3850220
HU	Hungarian	277518	9865769
LV	Latvian	67736	1342180
NO	Norwegian	405039	8938168
NL	Dutch	877590	24881686
PL	Polish	903670	29731309
PT	Portuguese	1088962	24867864
RO	Romanian	307084	5392290
RU	Russian	1232353	27442593
SK	Slovak	173417	4873409
SL	Slovenian	146250	5236834
SR	Serbian	239904	5013264
SV	Swedish	623035	11515290
TR	Turk	304853	9557808
UK	Ukrainian	322799	9158661
ZH	Chinese	609262	15838584

Table 1: Basic graph properties of analysed corpora and their corresponding ISO639-1 codes

For every corpus all contained page titles were ordered according to their descending PageRank values. We call such a list to be an *intracultural list* and we call *langrank* the placement of a given item in its respective *intracultural list*. Hence, 27 *intracultural lists* were obtained within which pages have *langrank* 1, pages with second highest probabilities have *langrank* 2, etc. To summarize, high langrank means low PageRank importance and vice versa.

To detect what names of countries are to be found on the very top of intracultural lists (i.e. have lowest langrank), a following procedure was applied: a term with langrank position 1 was extracted from the list, and translated it into English by using wikipedia itself as the translator. If it was not present in the ISO list of country names, procedure continued with a term having langrank position 2, 3, etc. If it was in the list, the procedure continued with country detection in following intracultural list, therefore repeating itself 27 times.

2.2. Results

27 intracultural PageRank vectors, one for each language community, were obtained and subsequently ordered in descending order according to calculated PageRank (converged probability) value. For illustration, in Tab. 2 we offer «top 10» values of such lists for 2 Latin and 2 Slavic corpora.

Portuguese		*Spanish*		*Czech*		*Russian*	
Wikipédia	0.065305	**España/Sección**	0.491755	Wikipedie	0.00984	Википедия:Справка	0.01519
Proxy	0.006393	Rural	0.050179	Wikimedia_Commons	0.00816	**Русская_Википедия**	0.00564
WP:TT	0.003323	Wikipedia	0.001105	GNU_Free_Documen-	0.00303	Германия	0.00361
Plantae	0.002419	Wikipedia_	0.000887	tation_License|		Общественное_достояние	0.00348
Til	0.001981	en_español		CC-BY-SA	0.00141	GNU_Free_Documentation	0.00295
Avaré	0.001496	2001	0.000555	CAPTCHA	0.00132	_License	
População	0.001492	Mayo	0.000508	**Česko**	0.00109	Викисклад	0.00277
Invertebrados	0.001435	Wikimedia_	0.000337	IP_adresa	0.00097	Creative_Commons	0.00276
Área	0.001433	Commons		Spojené_státy_americké	0.00082	Английский_язык	0.00121
Brasil	0.001412	GFDL	0.000205	Zeměpisné_souřadnice	0.00079	**Россия**	0.00119
		España	0.000197	**Praha**	0.00069	Фонд_свободного_програ	0.00112
		Rural	0.000196				

Table 2: Top ten (i.e. langrank 1 – 10) items of 4 intracultural lists and their respective PageRanks

It may be easily observed from the data that Wikipedia itself holds one of the top positions (this is the case within other 23 corpora as well). This is a trivial discovery since a wikipedia system is designed in the way that it refers in the first place to articles which concern the functioning of the system itself. Slightly less trivial is the observation that articles concerning the names of countries or cities closely associated to a language of a given wikipedia corpus emerge at the top positions of their respective intracultural lists.

Wiki corpus	*Top country*	*L*	*Wiki corpus*	*Top country*	*L*	*Wiki corpus*	*Top country*	*L*
AR	مصر (Egypt)	17	**FR**	France (France)	23	**RO**	România (Romania)	7
BG	България (Bulgarria)	4	**HE**	ישראל: (Israel)	7	**RU**	Германия (Germany)	3
CS	Česko (Czech Republic)	6	**HR**	Hrvatska (Croatia)	4	**SK**	Slovensko (Slovakia)	9
DA	Danmark (Denmark)	34	**HU**	Magyarország (Hungary)	18	**SL**	Slovenija (Slovenia)	8
DE	Deutschland (Germany)	16	**LV**	Latvija (Latvia)	6	**SR**	Француска (France)	28
EL	Ελλάδα (Greece)	7	**NL**	Frankrijk (France)	11	**SV**	USA	35
ES	España (Spain)	9	**NO**	Norge (Norway)	6	**TR**	Türkiye (Turkey)	13
ET	Eesti (Estonia)	5	**PL**	Polska (Poland)	12	**UK**	Україна (Ukraine)	13
FI	Suomi (Finland)	5	**PT**	Brasil (Brazil)	10	**ZH**	印度尼西亚 (Indonesia)	10

Table 3: Country names found at the top of their intracultural lists (i.e. having lowest langrank L)

Answers to the question «What countries are the first to occur at the top of given corpus intracultural importance list?» are present in Tab. 3. In 22 cases did an extraction of one country name from the top of the intracultural list corresponding to the graph of wikipedia written in language X yield the name of a country where this very language X is an official language of the state. Five exceptions are: Dutch where *Frankrijk* (L=11) closely outran *Nederland* (L=14); Russian where *Германия* (L=3!!!) outran *Россия* (L=9); Serb where *Француска* (L=28) far outran *Србија* (L=70); Swedish where *USA* (L=35) closely outran *Sverige* (L=37) and finally Chinese where Indonesia (L=10) is followed by Qatar (L=45), Micronesia (L=371), Brunei (L=409), Taiwan (L=484) and only much later by mainland China 中国 (L=579).

2.3. Discussion

The observation that *huge majority (22 out of 27) corpora yields in the top positions of their respective langrank lists the names of countries whose official language is identic to the language of corpora under study* is the first indication that even a pure hyperlink analysis could possibly reveal itself as a fruitful method for obtaining an overall information about preferences or interests of authors of wikipedia corpora. In such a manner could it posssibly serve as a means for «cultural stylometry» – a technique which could possibly allow to determine an appartenance of an anonymous author (or group of authors) to a given cultural or social unit.

For instance, data from Tab. 3 indicates that «central country of interest» for auhors of *PT* corpus is *Brasil* (L=10) and not Portugal which emerges only later in the list (L=32), later than *França* (L=12), *Itália* (L=14), *Espahna* (L=16) and even *Estados Unidos* (L=31). If a basic hypothesis of this article, i.e. that langrank values represent the amount of importance of a given term in a given corpus will not be falsified, it could be proposed that Brasil plays, for authors of PT corpus, much more important role than Portugal, from which it could be inferred that majority of them is possibly from Brazil and not from Portugal. Analogic stylometric conclusions can be inferred when looking at the *AR* corpus where Egypt (L=17) is followed by Jordan (L=27), Spain (L=36), France (L=37) and Tunisia (L=47).

An interesting exception occurs for the countries for which the official language is not identical to the language of a country in which a wiki corpus was written: the fact that Netherlands is closely overran by France in case of Dutch corpus and Sweden by USA in case of Swedish corpus can be possibly interpreted by the proposing that the overall global currents – related more closely to cultural superpowers are, for wikipedia authors of these two highly developed nations, of slightly more interest than local current of nationalist nature.

The results obtained for Chinese intracultural list are intriguing. While a position of Indonesia of the very top could be naively explained by activity of Chinese expats in Jakarta who pass there time writing wikipedia articles, the subsequent emergence of Qatar, Micronesia and Brunei seem to be completely contraintuitive. These phenomena can be, however, explained by a well-known caveat of PageRank algorithms related to so-called *linksink* phenomenon. A linksink can emerge during the PageRank vector calculation when the analyzed graph contains a densely interconnected subgraph having only few links to the rest of the graph. One way how to deal with linksink perturbations is an optimization of damping factor, these problems in relation to our cultural comparative method will be addressed in following articles.

Since the top of Serbian intracultural list indicates that this corpora is subject to linksink perturbations (first 45 positions are occupied solely by astronomic terms), we consider this to be an explanation for the observation where Serbia is far overran by France. Since Serb corpus is not a big one, the result can be as well explained by an overly activity of a small group of authors biased more towards France related phenomena than to Serb related ones.

Striking fact that Germany occupies third position in Russian intracultural importance list is left for reader's interpretation.

3. "The world&corpus" study

While huge majority of results obtained during analysis 1 seem to be consistent with intuitive expectations, their true scientific significance remains discutable. To address this issue, we have conceived a second analysis in which we have decided to correlate precalculated intracultural lists with factual data. For this purpose we have decided to use the real geographic (spatial)

distances between the country of a linguistic community under study, and other country (i.e. country of reference). Such a choice was motivated by a simple hypothesis: wikipedia users from *home country* B will, more likely, write articles and create hyperlinks concerning countries of reference A and C which are neighbours of B, than about countries of reference X or Y which are spatially distant. If such a tendency exists, and if PageRank is a sufficiently efficient technique for quantification of such an "importance" of A, C, X, Y countries of reference within the scope of corpus created by authors supposedly from home country B, then significant correlations between intracultural lists and |home country, country of reference| spatial distance can be expected to occur.

3.1. Method

We have defined 32 *countries of reference*: 27 of them were countries which we have considered as well to be *home countries* of our intracultural lists; 5 others were chosen by random, one from every continent (Italy, Japan, Senegal, Argentina, Australia).

As a first dataset we have used 27 intracultural lists, one for each *home country*, calculated during analysis 1. From every such list, the langrank (i.e. position sorted according the ascending pagerank value) corresponding to the the term denoting the country of reference was extracted. For example, as Tab. 4 illustrates, Hrvatska was on the 4th position in a Croatian corpus and 74th in Slovenian corpus.

Language of home country	*Langrank position*	*Name of country of reference*	*Spatial distance (km)*
AR	532	كرواتيا	3464
BG	345	Хърватия	797
CS	281	Chorvatsko	509
DA	848	Kroatien	1265
DE	329	Kroatien	808
EL	271	Κροατία	870
ES	756	Croacia	1695
FI	456	Kroatia	2197
FR	1131	Croatie	1056
HE	1493	קרואטיה	2255
HR	4	Hrvatska	0
HU	268	Horvátország	403
LV	675	Horvātija	1472
NL	409	Kroatië	1083
NO	418	Kroatia	1907
PL	422	Chorwacja	828
PT	749	Croácia	2028
RO	469	Croaţia	746
RU	696	Хорватия	5533
SK	271	Chorvátsko	494
SL	74	Hrvaška	118
SR	110	Хрватска	455
SV	556	Kroatien	1874
TR	413	Hırvatistan	1747
UK	679	Хорватія	1320
ZH	3981	克罗地亚	7321

Table 4: positions of country of reference Croatia in intracultural lists of diverse home countries and their spatial respective distance

Mathematica functions of computational search engine «Wolfram Alpha» were used as a resource of *home country* ↔ *country of reference* spatial distance data.

Pearson correlation coefficients were calculated between two datasets. Whole procedure was repeated 32 times, once for every country of reference.

3.2. Results

Obtained results suggest significative correlations between intracultural lists and geographic data in case of all countries of reference with exception of China, Russia and Slovakia. They are presented in Tab. 5.

3.3. Discussion

Obtained results show correlations between strongly empiric spatial measures and positions within the "intracultural" lists Since different wikipedia corpora are direct consequences of different creative preferences of human groups, these correlations have to be explained in terms of these preferences. We propose that these preferences are culturally determined.

The previous analysis even if it leads us to interesting conclusion, is however questionable. And a major caveat should be raised: Pearson's correlation coefficients are sensitive to outlier datapoints and if these are present, an analysis cannot be considered as a robust one (Rousseeuw and Leroy, 2003).

Country of ref.	*p*	*cor*	*Country of ref.*	*p*	*cor*	*Country of ref.*	*p*	*cor*	*Country of ref.*	*p*	*cor*
Argentina	<0.003	0.549	Finland	<1.74E-05	0.727	Latvia	<5.6E-05	0.696	Senegal	<0.0007	0.617
Australia	0.165	-0.275	France	0.0015	0.577	Netherlands	<0.007	0.507	Slovakia	0.1965	0.256
Bulgaria	<0.00026	0.648	Germany	<0.004	0.539	Norway	<0.0003	0.652	Slovenia	<6.63E-07	0.797
Croatia	<2E-06	0.779	Greece	0.00019	0.657	Poland	<0.0005	0.630	Serbia	<9.53E-05	0.680
China	0.426	0.183	Hungary	0.00015	0.664	Portugal	<0.05	0.387	Spain	<0.011	0.486
Czech R.	<7-E05	0.689	Israel	0.0148	0.463	Romania	<6.8E-05	0.690	Sweden	<0.001	0.599
Denmark	<0.00044	0.629	Italy	<0.005	0.525	Russia	0.8987	0.025	Turkey	<0.0004	0.635
Estonia	<1.5E-05	0.730	Japan	0.711	-0.07	S.Arabia	<0.0035	0.543	Ukraine	<0.0005	0.629

Table 5: Overall p-values and Pearson correlation coefficients (d=25) for 32 countries of reference

As Fig. 1 illustrates, this was the case for example in the situation when Germany was chosen as a country of reference. Simple removal of *zh* (Chinese) datapoint from the top right corner (i.e. high spatial distance, high langrank) have caused a drastic change from (cor=0.539; p<0.004) to (cor=-0.108; p=0.599). Since majority of countries of references in analysis 2 were European ones, it can be expected that this outlier boosts up the significativity of our hypotheses in an unwanted manner.

Another source of bias was identified as well. It is related to the fact that Wolfram Alpha uses cartographic center of a country as the point from which it measures a distance to/from a given country. That's a useful feature in case of countries whose population is distributed equally. In case of a country like Russia, however, is the *ru* "central point" postulated somewhere in central Siberia, 4000 km east from Moscow. Whether such a point can have anything to do with cultural preferences of wikipedia authors is a place for argument.

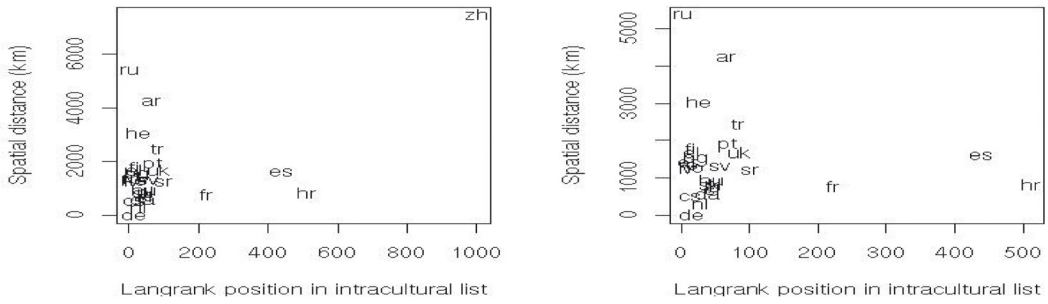

Figure 1: Visualisation of langrank&distance correlations when « China » outlier is included (left) in or excluded (right) from the list of countries of reference as related to Germany

4. General Discussion

The aim of "the top country" study was to demonstrate whether a method of parallel pageranking of wikipedia graphs can yield relevant information concerning basic overall specificities of the corpora, and therefore of their authors. Simple look up at the tops of calculated intracultural lists have demonstrated that such is verily the case: in 22 out of 27 corpora was the topmost ranked country-concerning article about the country whose official language is that in which the corpus was produced.

The second, "world&corpus" study focused on a relation between implicit properties of wikipedia corpora and geographic distances of the factual world. While significativity of obtained results suggest that there possibly exist some morphic relations between the overall hyperlink structure of (wikipedia) corpora and the factual world, the outlier problem indicates that the "world&corpus dilemma" will not be an easy dilemma to resolve.

What we denote here as "world&corpus dilemma" is only very superficially related to method which we presented in our second study. In fact, it is much more closely related to an ancient epistemological problem "What is knowledge and how is it represented?" than to some trivial linear regression of two sets of datapoints which tend to show to have something in common.

In its weaker form, the question goes like this "What is relation between the corpus and the world, given that corpus is sufficiently big?". The goal of our article was to indicate that the graph theory could possibly bestow a temporary question to this answer: "If a graph of the corpus is isomorphic with the graph of a world the corpus tends to describe, than it can be said that such a corpus contains the knowledge about that world".

We say "a" graph, because there are infinitely many ways how to construct a graph from a given corpus. For the purposes of this article, we have chosen the most simple way: inspired by "random surfer model", we have completely ignored information IN the Net (*e.g.* word co-occurences in the content) and focalized at the information ON the Net.

An edge have been created when a hyperlink existed between the vertices. We supposed this assumption should be suffice as a *point de depart: the very act of creation of an article, or a hyperlink, can be an interesting clue to the preferences of the one who creates it*. A weak clue, of course, but nonetheless containing more information than pure accident.

Since it is well known that a well aggregated linear combination of weak classifiers can result in a highly-effective strong classifier (Freund and Schapire, 1996), it can be as well proposed that a huge number of well aggregated weak cultural clues can yield some strong ones.

References

Adar E., Skinner M. and Weld D.S. (2009). Information arbitrage across multi-lingual Wikipedia. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, ACM, pp. 94-103.

Bourdieu P. (1979). *La distinction: critique sociale du jugement*. Paris: Ed. de Minuit.

Brin S. and Page L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer networks and ISDN systems*, 30 (1-7): 107-117.

Csárdi G. and Nepusz T. (2006). The igraph software package for complex network research. *InterJournal Complex Systems*, 1695.

Esuli A. and Sebastiani F. (2007). PageRanking WordNet synsets: An application to opinion mining. In *Annual meeting-association for computational linguistics*. pp. 424-431.

Ferrandez S., Muñoz R. and Palomar M. (2007). Applying Wikipedia's multilingual knowledge to cross-lingual question answering. *Lecture Notes in Computer Science*, 4592, pp. 352-363.

Filatova E. (2009). Directions for exploiting asymmetries in multilingual Wikipedia. In *Proceedings of the Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies*, Association for Computational Linguistics, pp. 30-37.

Freund Y. and Schapire R.E. (1996). Experiments with a new boosting algorithm. In *Machine learning-international workshop then conference*, Citeseer, pp. 148-156.

Mihalcea R. (2007). Using wikipedia for automatic word sense disambiguation. In *Proceedings of NAACL 2007 HLT*.

Park H. (2005). Network Cultural Analysis: Texts, Graphs, and Tools. In *Paper presented at the annual meeting of the American Sociological Association*, Philadelphia, PA.

Richman A.E. and Schone P. (2008). Mining wiki resources for multilingual named entity recognition. *Association for Computational Linguistics (ACL-08: HLT)*: 1-9.

Rousseeuw P.J. and Leroy A.M. (2003). *Robust Regression and Outlier Detection*. Hoboken, New Jersey : J. Wiley & Sons.

Surowiecki J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. New York: Doubleday Books.