Maps, spaces, and phylogenetic trees of the Internet governance ecosystem

~ Preliminary report ~

Goran S. Milovanović,

DiploFoundation, Belgrade

The four days of the 9th Annual Internet Governance Forum (IGF), held in Istanbul, Turkey, 2 - 5 September 2014, coincided with interesting developments in Diplo's methodological approach to Internet governance (IG). The IGF provided us with the opportunity to study the statistical properties of the emerging language of Internet Diplomacy as it unfolded before us in numerous workshops and sessions. Our experiments and exercises were conducted in the scope of the development of Diplo Text-Analytics Framework (DTAF). This analytical framework utilises text-mining technology in order to foster the interpretation of immense amounts of textual input to global multistakeholder governance processes. As the experience of NETmundial witnesses, the ability to develop an understanding of a large number of technical and non-technical concepts - some of them introduced in the course of debate - map the relations between them, and coordinate the statements of numerous stakeholders in an efficient way, may soon prove to be a capacity of indispensable value in the management of complex governance processes.

'For a large class of cases of the employment of the word "meaning"- though not for all - this can be explained in this way: the meaning of a word is its use in the language.'[i] This sentence, found in Ludwig Wittgenstein's *Philosophical Investigations* (originally published in 1953), presents a cornerstone of the *use theory of meaning*, in the development of which the great philosopher partially abandoned his previous 'pictorial theory' of meaning, switching from a correspondence-based towards a more relational, context-based understanding of how the concept of meaning works in human language and thought. In the following text, we will pursue both perspectives: we will show how a search for the way words are used unveils their meaning across diverse contexts, and we will literally *look* for the meaning of words[ii] in maps and graphs that depict the evolving semantics of the IG ecosystem. As the analogy that maps the concept of IG onto the concept of ecosystem gained significant strength in the course of the contemporary debate, we will foster the use of biological metaphors in the presentation of our results, in order to better align with the current context.

To begin with, the plot in Figure 1 - called a *Hillis plot*, to honor the evolutionary biologist David Hillis[iii] who invented it to depict the evolutionary tree of life - maps a structure of similarities between more than 100 words and phrases that occurred during the 9th IGF in Istanbul:

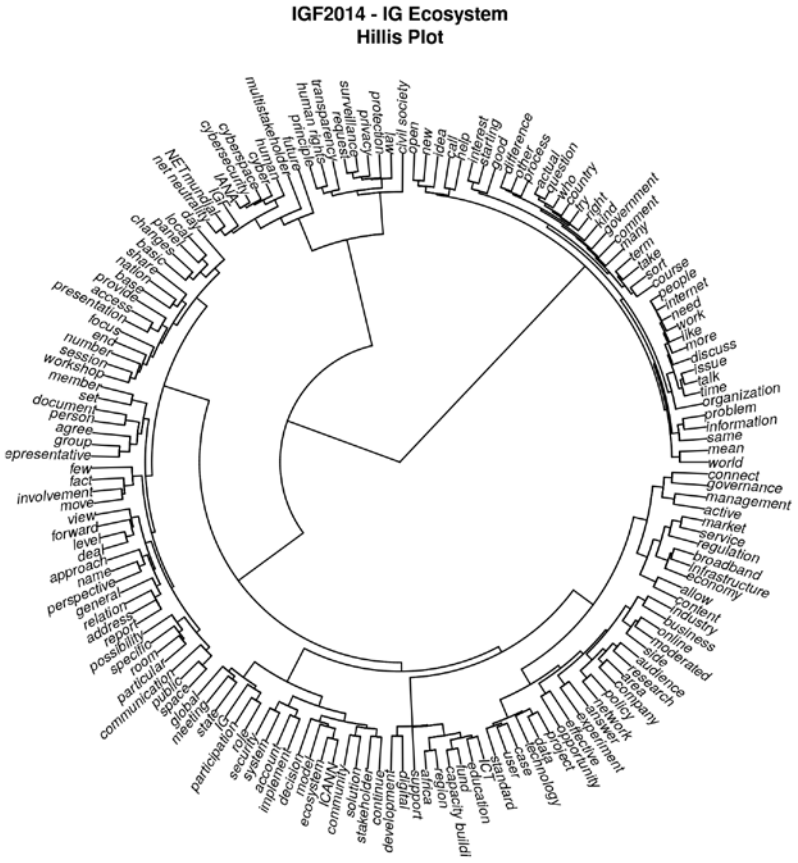

Figure 1. The IG ecosystem as presented in the IGF 2014 Text Corpus: Hillis plot.

Look at approximately 11 o'clock where you can find *NETmundial* and *net neutrality* adjacent to each other. They are grouped under one single branch of the Hillis plot, a branch that first proceeds to encompass *IGF*, than join with the group that encompasses the concepts of *IANA*, *cybersecurity*, *cyberspace*, *cyber*, and *human*; the whole branching discussed thus far is then joined by *multistakeholder* and *future* at the next level of branching. We can start at different positions at the exterior of the plot and study the way words and phrases progressively group towards the interior of the plot according to how similarly they were used during the IGF. Finally, all words and phrases converge towards the center of the Hillis plot, at the origin of the most abstract level of categorisation, which we might call simply *The IG discourse*. We see that at the most abstract level of categorisation, in the centre of the plot, two branches kick-off towards its exterior, of which the left one soon splits in two; by following the plot from the most abstract level we can recognise the most robust grouping of words and phrases - the most extensive and abstract categories of concepts in the IG discourse

But how do we detect the similarity in use of different words and phrases? Where do they come from and how are they processed before such structures as the Hillis plot emerge? Here is the story.

In the beginning, we have at our disposal a number of documents, encompassing a number of utterances produced by various stakeholders that express their positions in reference to different IG-related issues. Precisely, we start with the collection of 91 raw IGF 2014 session transcripts.

How do we get from a pure collection of documents to networks of words and phrases that map the discussion? In a nutshell, here is what we do. We study the distribution of word frequency across various documents. Imagine starting with three text documents only. We count the number of occurrences of the word *Internet* in these three texts and the result is: 15 (in the first document), 4 (second document), 17 (third document). Assume that we next count the occurrences of the words *gallery* and *governance* across the same documents, and find the frequency distribution of 13, 2, and 18 for *governance,* and 2, 19, and 1 for *gallery*. Interestingly, the usage of words *Internet* and *governance* is related across the documents: they are both frequently used in the first (15, 13) as well as in the third text (17, 18), while rather rarely in the second document (4, 2). The usage pattern for *gallery* is rather opposite: it is frequently used in the second document (19), and rarely used in the first (2) and the third (1). Our conclusion is two-fold. First, the words *Internet* and *governance* are somehow related, since the authors of different documents tended to increase the use of one of them each time they increased the use the other, and by following the same logic we know that the use of the word *gallery* is less related to the way *Internet* and *governance* were used. Second, we have learned something about the nature of the documents: the first and the third text have probably something to do with Internet governance, while the second one has probably something to do with running galleries, opening exhibitions, and enjoying art.

Of course, we don't want to base our conclusions on the study of three words in three documents. That is why we have hand-picked an IG- specific terminological model, encompassing more than 5000 words and phrases whose frequency distributions undergo analysis. We collated the complete transcripts of the IGF 2014 sessions in the text corpus on which we base our observations. Everything beyond this is based on the application of computational procedures and mathematical statistics. As we were able to recognise the similarity in the use of two words in three documents, computers help us discover the similarity network of thousands of words and phrases across many documents. Mathematical models are used to *translate these similarities into distances*, so that we place words and phrases on *maps* and *high-dimensional spaces*, where similar concepts tend to live together, as similar climates tend to be grouped across world regions or similar plants and animals inhabit similar ecosystems. By even more advanced procedures, we turn these complicated maps and spaces into *graphs* to explore the *semantic neighbourhoods*: groups of words related by their similar use across collections of documents.

For the following analysis, we selected 31 IG-specific words and phrases. We extracted their frequency distributions across 91 IGF sessions, and then computed the distances between them[iv] in order to be able to represent them spatially. The initial spatial representation is high-dimensional and thus not of much use for the human visual system. We then applied a dimensionality-reduction technique[v] to represent these words and phrases in 2D and 3D spaces (a 2D space is really a map where all points lie on a plane). Figure 2 presents a 3D semantic space of the selected words and phrases. The more similarly the two concepts (words, phrases) were used across 91 sessions, the closer they stand in the semantic space; in other words, spatial distance corresponds to *dissimilarity*. A similar approach led to the production of the Hillis plot in Figure 1.[vi]

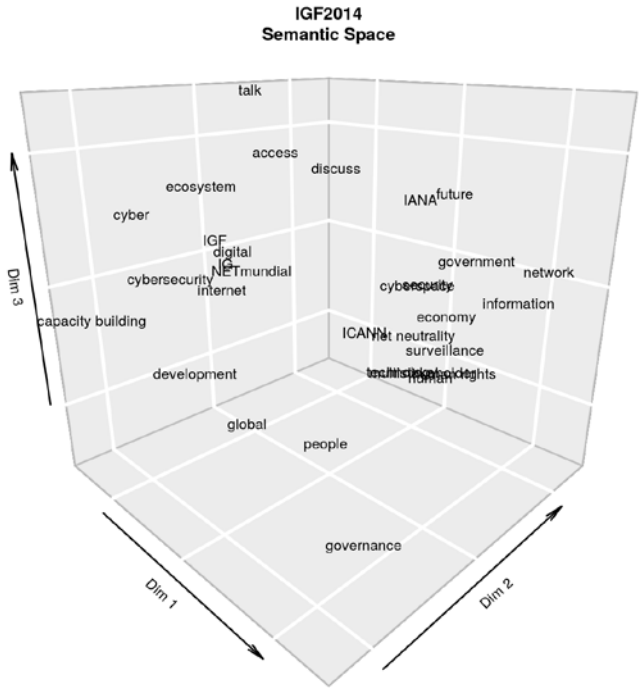

Figure 2. Selected IG-specific words and phrases in a 3D semantic space.

However, semantic spaces can be difficult to comprehend, even in bearable three dimensions dear to a visual system like ours. To ease the interpretation of the semantic space, we introduce the following simple concept of *N-neighborhood*: N concepts most proximal to a particular concept comprise that concept's N-neighborhood. We choose N = 3 and produce a graph[vii] in Figure 3 that depicts the 3-neighbourhoods for all selected words and phrases under analysis. In Figure 3, all words and phrases are represented by nodes on the graph. Each node receives exactly three inputs: these are the arrows that point towards that node. The three nodes that extended their arrows towards a particular node comprise that node's 3-neighbourhood. Now we can study the pattern of connectivity that holds between the IG-specific concepts: try and find out whether there is a way to connect any pair of concepts presented in Figure 3.

Start from *ICANN* and you will easily find your way to *IANA*. However, try to bridge *ICANN* or *IANA* with *NETmundial* or *IGF* and see what happens. On the left, we find a large pattern of connectivity extending over more **strategic** concepts: *IG*, *IGF*, *NETmundial*, *Internet*, *ecosystem*, *capacity building*, *development*, *cybersecurity*, etc. On the right of the graph, approximately delineated by the line that connects *access* and *digital*, we find a pattern of connectivity extending over more **implementation** concepts: *ICANN* and *IANA* are found there, as well as *multistakeholder*, *government*, *economy*, *net neutrality*, *human rights*, *governance*, *technology*, *surveillance*, *security*, etc.

Of course, we have had to reduce the true complexity of the semantic space in order to produce this analytically tractable representation. Different choices of words and phrases that undergo analysis, the selection of broader N-neighborhoods, as well as the selection of more specific document sets upon which the observations are made, would all produce different semantic spaces and different patterns of connectivity. We should think about these various choices as **switching perspectives** or **viewpoints** from which we choose to observe the way words and phrases group and connect.

Figure 3. Semantic 3-neighbourhoods of selected words and phrases.

Figure 3 thus presents a rather rough approximation of the true complexity of the discourse. However, as the true complexity surpasses any limits of interpretation that are useful to our minds, we are building a text-analytics framework that lets us choose any arbitrary level of precision in the dissection of the IG discourse. Our text-analytics platform is being developed to help IG actors figure out the possible perspectives and where the main lines of the debate fall. It was not built to (and, more important, *it should not*) think instead of us. Thus, after the computing machinery finishes its job, interpretation remains an art of finding the right balance between simplicity and exactness.

Figure 4 presents semantic neighborhoods of selected IG-specific terms following the change of perspective in respect to Figure 3. We have removed several words and phrases and introduced two new concepts: *education* and *science*. As readily seen from Figure 4, *science* naturally finds its way in the neighborhood of *technology*, *information*, and *multistakeholder*, while *education* joins *capacity building*, *Internet*, and *development* on the right. The neighborhood of *ICANN* now encompasses *science*, *human*, and *governance*, while that of *IANA* comprises *economy*, *network*, and *information*, in contrast to *surveillance*, *government*, and *economy* in Figure 3. Note how *IGF*, *IG*, and *NETmundial* remain fully connected in both Figure 3 and Figure 4, irrespective of the changes in the set of words and phrases under analysis.

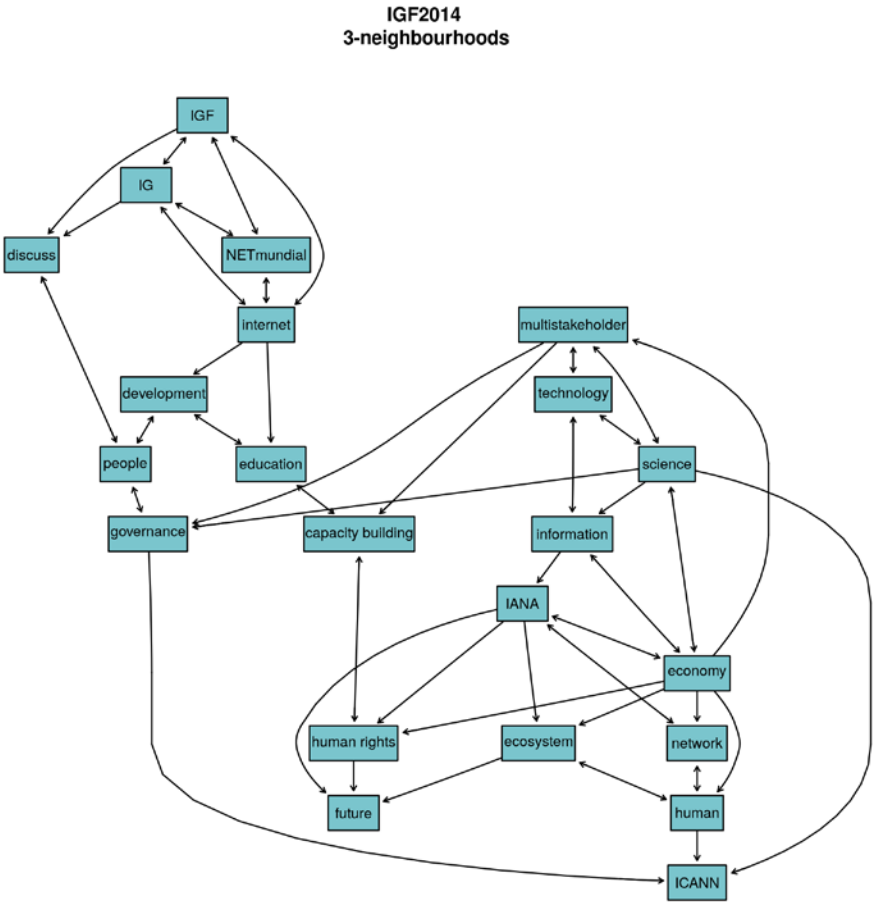

Figure 4. Semantic 3-neighbourhoods of another set of selected words and phrases.

Now we have seen an example of the effect of changing perspectives. Complex discussions call for complex tools to analyse them; anyone who has ever tried to study IG has learned that there is no easy way through the forests and labyrinths of the contemporary IG debate. In DiploFoundation, we are building tools that might help downsize this complexity to manageable proportions, and yet enable precise insights into specific details when necessary. The current version of our Text-Analytics Framework was developed in less than one month before the onset of IGF 2014; the development of more advanced features is underway.

Notes

Text-mining procedures were developed in R, using and extending the functionality of the *tm*() package[viii] for text preprocessing and the production of term-document matrices.

The R Project for Statistical Computing
http://www.r-project.org/

tm: Text Mining Package
http://cran.r-project.org/web/packages/tm/index.html

This preliminary report appeared on DiploFoundation's website, published as a blog post, in two parts (12/15 September 2014).

---

[i] Wittgenstein L (1953/2001) *Philosophical Investigations*. Blackwell Publishing: Oxford, UK.

[ii] Our pictures, however, are not of a kind envisioned in Wittgenstein's early pictorial theory. While in his early works they stand for basic states of affairs of atomic facts, in ours they map the relations of words between themselves and across documents, essentially mapping language onto itself. Thus we are almost offering a diplomatic compromise to the thoughts of the great philosopher here.

[iii] David Hillis (1958 -), http://en.wikipedia.org/wiki/David_Hillis

[iv] *Variation of information metric* was used to represent the distances. Given the co-occurrence vectors for any two concepts, the function $d(x,y) = H(x) + H(y) - 2I(x,y)$, where $H(x)$ and $H(y)$ are the empirical Shannon's entropies of the vectors x and y, respectively, and $I(x,y)$ is the empirical mutual information, satisfies the metric space axioms.

[v] *SMACOF* package in R (De Leeuw J and Mair P (2009) Multidimensional Scaling Using Majorization: SMACOF in R. *Journal of Statistical Software*, *31*(3), 1–30. Available at http://www.jstatsoft.org/v31/i03/paper; technical documentation is available from *SMACOF: multidimensional scaling in R, available at:*http://cran.r-project.org/web/packages/smacof/index.html) was used to perform ordinal multidimensional scaling (MDS). Ordinal MDS solutions are invariant up to monotonic transformations, i.e., they try to preserve the rank-order of original distances from original high-dimensional spaces. For the MDS configuration in Figure 2, *Stress* = .15; for the 3D MDS configuration from which the graph in Figure 4 was produced, *Stress* = .15 also.

[vi] Hierarchical cluster analysis was performed by *hclust*() in R. The Hillis plot was produced from the *ape* R package for analysis of phylogenetics and evolution (The ape package is available at http://cran.r-project.org/web/packages/ape/ape.pdf).

[vii] *RGraphviz* package (available at http://bioconductor.wustl.edu/bioc/html/Rgraphviz.html) was used to produce the 3-neighbourhood graphs, previously instantiated from incidence matrices using the basic functionality of the R *graph*() package (available at http://www.bioconductor.org/packages/release/bioc/html/graph.html).

[viii] Feinerer, I., Hornik, H. & Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software*, Vol. 25, Issue 5, Mar 2008. URL: http://www.jstatsoft.org/v25/i05/paper