Contribution ID: **6** Type: **not specified**

Relation between Zipf ’s law and the distribution of shared components in complex component systems.

Tuesday, 27 September 2016 15:20 (20 minutes)

Several complex systems of diverse nature consist of realizations which can be broken into their elementary constitutive components, for example, books into words, genomes into genes, and technological systems into building blocks.
The statistics of the components (e.g., word) across realizations (e.g., books) shows several quantitative laws, such ad the well-known example of the power-law distribution of component abundances, known as Zipf's law in the context of natural languages.
Central to the current debate in evolutionary genomics is a different law, the ”gene-frequency distribution”, or ”occurrence distribution”, where a component
occurrence is defined as the fraction of realizations in which the component is present.
In genomes, the occurrence distribution shows a peculiar ?U-shape? due to a large number of rare (i.e. belonging to very few species) and common genes (present in almost all the species), compared to genes at intermediate occurrences.
While several possible theoretical explanations of the U-shaped gene occurrence distribution have been proposed, its causes are still under debate.
Here, we consider occurrence distributions in three datasets from genomics, linguistics (literary texts), and technology (LEGO toy constructions), showing that the U-shape is linked to the component frequency (i.e., the Zipf's law).
By means of a theoretical model based on sampling, we establish an analytical relationship between these two laws, which allows us to identify the crucial parameters affecting the occurrence distribution power law decay and the size of the common component peak.
The null model captures some relevant empirical features, as well as highlighting deviations that carry important information about the specificity of each system.

Primary authors: Dr MAZZOLINI, Andrea (Physics Department and INFN, University of Turin); Dr OSELLA, Matteo (Physics Department and INFN, University of Turin,)

Co-authors: Dr DE LAZZARI, Eleonora (Universite Pierre et Marie Curie, Paris, France); Dr COSENTINO LAGOMARSINO, Marco (Universite Pierre et Marie Curie, Paris, France); CASELLE, Michele (TO)

Presenter: Dr MAZZOLINI, Andrea (Physics Department and INFN, University of Turin)

Session Classification: Sessione 5