CERIAS

The Center for Education and Research in Information Assurance and Security
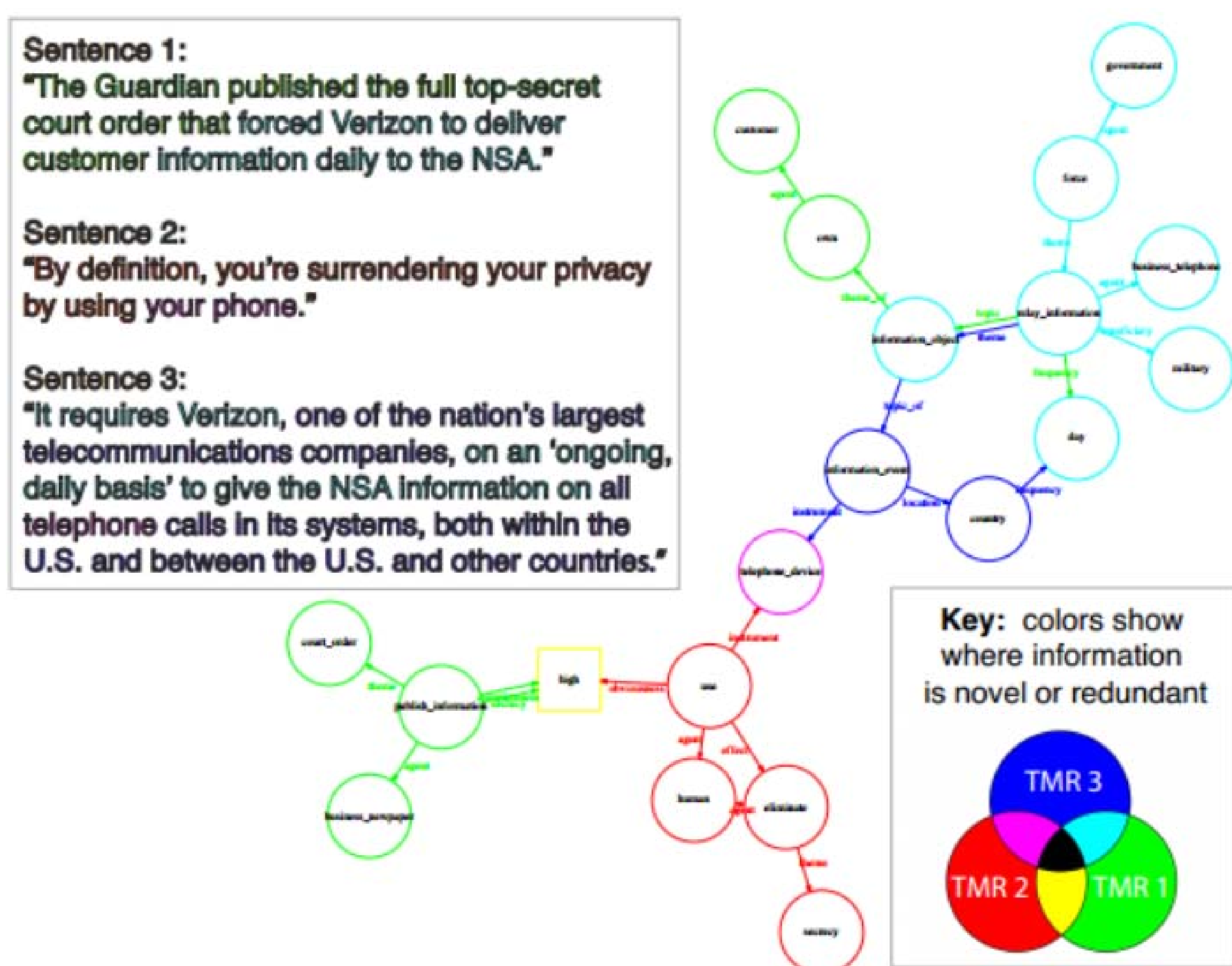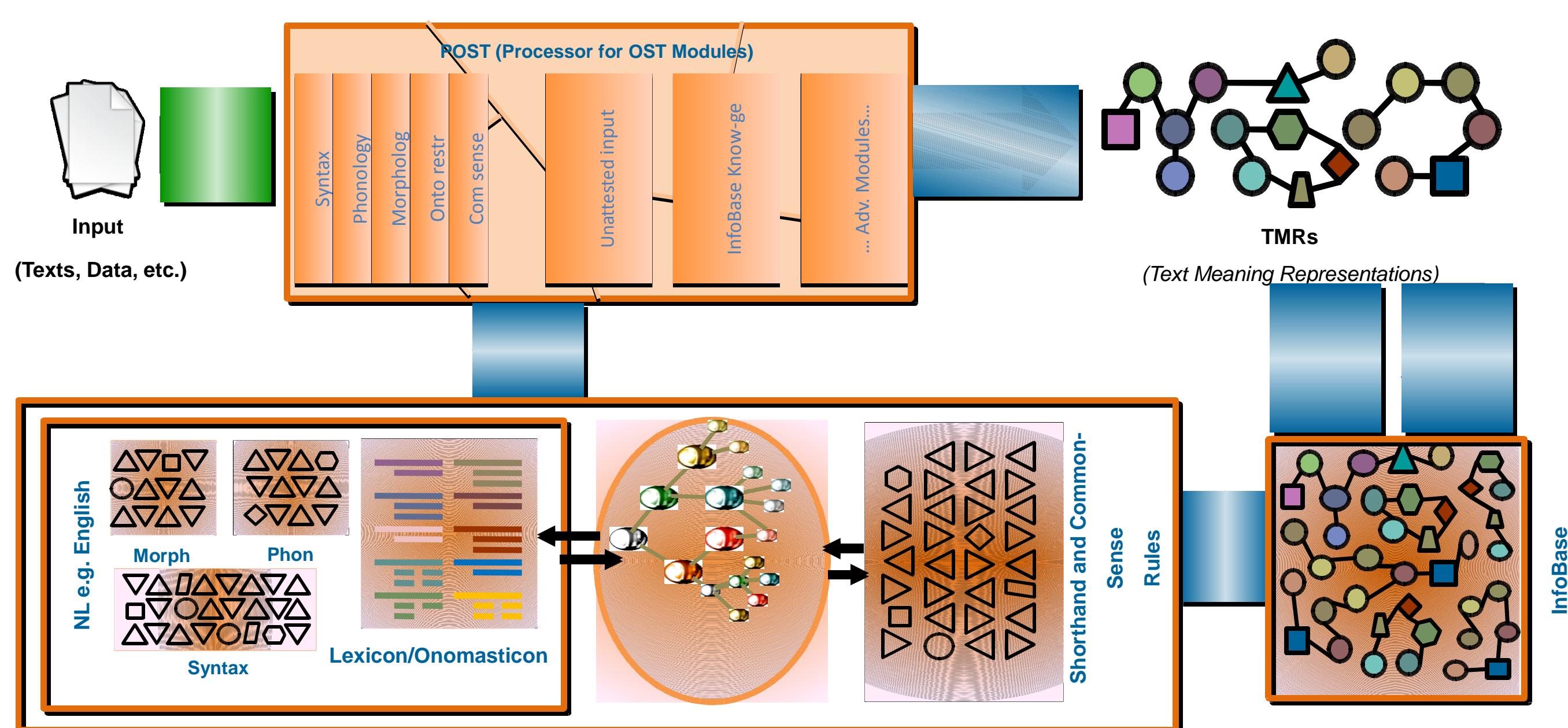
Semantic Anonymization of Medical Records

Student: Tatiana R. Ringenberg; Advisors: Julia M. Taylor & Victor Raskin

Overview

With the availability of large amounts of data in the medical industry, it is becoming necessary, due to both regulatory and ethical concerns, to find unique ways of protecting patient identities. A name and social security number are no longer the only fields in a patient's record that can identify them. Data under HIPAA requires the removal of several Protected Health Information Identifiers. Symptoms themselves can also distinctly identify an individual in a large group. To prevent this, the Purdue OST Anonymization Project is using semantics to determine the degree to which any patient record is identifiable from others in a system. Our approach combines the conceptual mapping of Ontological Semantic Technology with the anonymity principles of K-Anonymity to semantically anonymize patient data for compliance with regulatory and research policies.

Process

Ontological Semantic Technology (OST) is a theory, methodology and, especially, technology for representing natural language meaning, for automatic transposition of text into the formatted text-meaning representation (TMR), and for further manipulation of TMRs for inferencing and more advanced reasoning, both theoretically and in a growing variety of applications. The process of OST is described below.


Semantic Anonymization

Our approach for anonymizing patient medical records, using semantics, is based on the principles of K-Anonymity combined with OST.

K-Anonymity provides a means of generalizing data by altering fields, known as quasi-identifiers, which can be combined to identify an individual. These fields often include seemingly innocuous data such as zip code, age, race or sex. A set of records with K = 2 follows.

Age	Sex	Zip Code	Symptoms
20 - 30	f	4790*	Nausea
20 - 30	f	4790*	Headache
30 - 40	f	4790*	Cancer
30 - 40	f	4790*	Arthritis

Using OST, we generate a TMR to represent a patient's record. We then map that TMR with others, based on grain size, within an Infobase. Using K-Anonymity, we generalize the bottom most row of the concept map by either splitting the concept or moving it one step up the map. This is repeated until each entry is indistinguishable from K-1 other entries. The concept map is shown below.
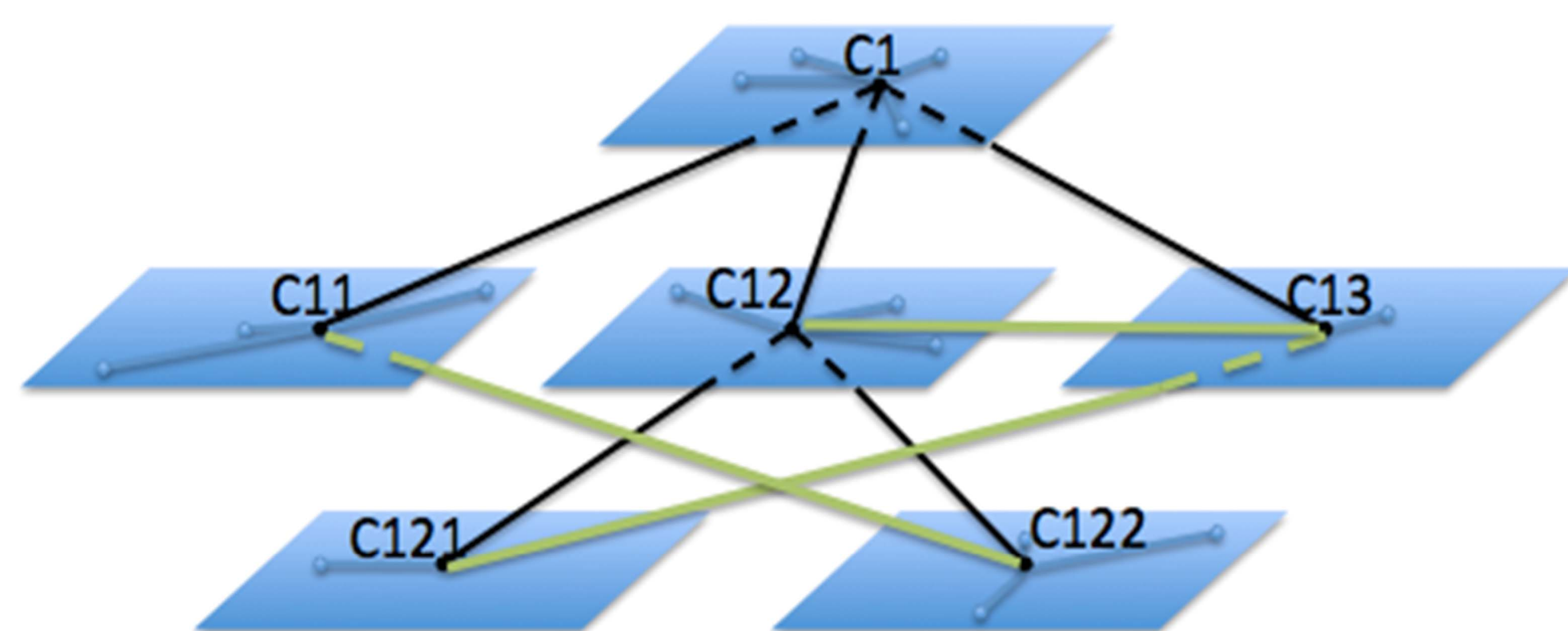
References

Graci, A., Stuart, L. M., Taylor, J. M., & Raskin, V. (2013). Towards novelty detection with ontological semantics technology. Proceedings of SDPS Conference 2013, São Paulo, Brazil.

Raskin, V., & Taylor, J. M. "The (not so) unbearable fuzziness of natural language: The ontological semantic way of computing with words." *Fuzzy Information Processing Society, 2009. NAFIPS 2009. Annual Meeting of the North American*. IEEE, 2009.

Taylor, J. M. (2012). *The use of natural language semantics in medical information*. Presentation Delivered At 3rd international workshop on ehealth data and knowledge, Kyung Hee University, South Korea.

PURDUE UNIVERSITY