Combinatorial Information Theoretical Measurement of the Semantic Significance of Semantic Graph Motifs

Cliff Joslyn*
Pacific Northwest National Laboratory
cjoslyn@pnl.gov

Sinan al-Saffar
Pacific Northwest National Laboratory
sinan.al-saffar@pnl.gov

David Haglin
Pacific Northwest National Laboratory
david.haglin@pnl.gov

Lawrence Holder
Washington State University
holder@wsu.edu

ABSTRACT

Given a semantic graph data set, perhaps one lacking in an explicit ontology, we wish to first identify its significant semantic structures, and then measure the extent of their significance. Casting a semantic graph dataset as an edge-labeled, directed graph, this task can be built on the ability to mine frequent *labeled* subgraphs in edge-labeled, directed graphs. We begin by considering the enumerative combinatorics of subgraph motif structures in edge-labeled directed graphs. We identify frequent labeled, directed subgraph motif patterns, and measure the significance of the resulting motifs by the information gain relative to the expected value of the motif based on the empirical frequency distribution of the link types which compose them, assuming independence. We illustrate on a small test graph, and discuss results obtained for small linear motifs (link type bigrams and trigrams) in the Billion Triple Challenge triplestore.

1. INTRODUCTION

As semantic graph databases (SGD) [10] grow, it is becoming increasingly important to be able to understand their inherent semantic structure, whether codified in explicit ontologies or not. Our research group is developing methods for **descriptive semantic analysis** of RDF triplestores, to serve purposes of analysis, interpretation, visualization, and optimization. We wish to identify the most prominent semantic structures and semantic constraints present in SGDs, first simply to understand them, but then to exploit them to provide targeted inferential support, and to optimize search and visualization methods to the specific ontology, connectivity, and distributional statistics of datasets and queries.

RDF[1] datasets are sets of triples $\langle s, p, o\rangle$, interpreted as both predicates $p(s, o)$ over the "resource" subject s and object o, and as graph links of type p from nodes s to o. Some predicates indicate semantic meta-data about resources, such as their classes $C(s), C(o)$. We have explored statistical representations of the structure of classes and predicates in semantic graph datasets [1, 8, 9], defining an **extant ontology** (EO) [8] as a class-predicate network over an entire RDF dataset, edge-weighted by predicate frequency. We also defined **ontological scaling** [1] as the ability to "roll up" classes and predicates through an external ontology to achieve coarser, more meaningful representations.

An EO is able to represent the *individual* link properties among node classes. However, the *joint* semantic constraints present amongst link types occurring *in combination* likely carries much more of the semantic information in a dataset. So we additionally explored the identification of significant path type structures as vectors of their constituent link types, basically link type n-grams [8, 9].

We now extend this work to address broader questions in graph data mining [4]. Methods in both network science and graph mining are aimed almost exclusively at unlabeled graphs, either directed or undirected [5, 11]. But semantics are exactly carried by the label information in the link types p and classes $C(s)$ and $C(o)$, in addition to the directionality of the links (triples are not generally symmetric). It may be valuable to know that two entities are connected by some path, but the exact nature of that path in terms of the intervening link types is critical. Similarly, query in graph databases is modeled as subgraph isomorphism down to matching the node and edge types of the query.

We aim to identify significant semantic structures by mining frequent *typed, directed* subgraphs as small motifs. We cast the typed link structure of an RDF dataset as an edge-labeled, directed graph, and define the combinatorial structure of its subgraph motifs. We use the SUBDUE program from Washington State University [3][2] to enumerate and count all such motifs. We then use an information gain measure, comparing the empirical frequency of (edge-labeled, directed) motifs to their expected frequencies based on the empirical distribution of their sub-motifs, assuming independence, down to the distribution of the individual edge labels. We illustrate on a small graph, and then show results for bigrams and trigrams of edge labels in paths of the

*Corresponding author: PNNL/Battelle Suite 400, 1100 Dexter Ave. N, Seattle, WA 98109 USA, +1-206-552-0351.

[1] http://www.w3.org/RDF

[2] http://ailab.wsu.edu/subdue

MDS 2011 Prepared for the Mining Data Semantics Workshop.

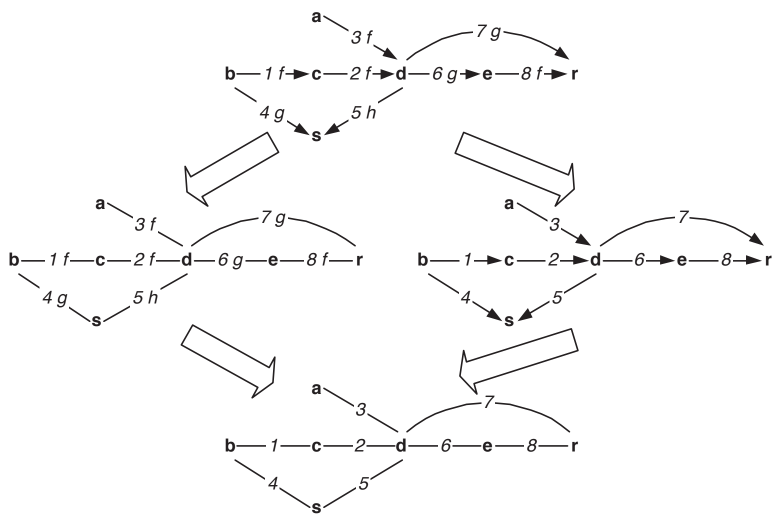

Figure 1: (Top) A labeled, directed graph. (Middle Left) It's undirected form, the symmetric closure. (Middle Right) Its unlabeled form. (Bottom) Unlabeled and undirected.

Billion Triple Challenge 2010 (BTC10) data set.[3]

2. ORDERED SETS OF MOTIFS IN LABELED, DIRECTED GRAPHS

We model an RDF triplestore as an edge-labeled, directed, connected graph $G = \langle V, E, \psi, L \rangle$, where: V is a finite, non-empty set of nodes, $E \subseteq V^2$ is a set of directed edges, L is a set of labels and $\psi \colon E \to L$ is a label function mapping each edge $\epsilon \in E$ to a label $\psi(\epsilon)$.[4] We will say that a graph has size $N = |E|$ with $\eta = |L|$ edge labels. An example is shown on the top of Fig. 1, with nodes $V = \{a, b, c, d, e, r, s\}$, $\eta = 3$ labels $L = \{f, g, h\}$, and each of the $N = |E| = 8$ edges $\epsilon \in E$ identified in $[[8]]$, where $[[x]] = \{1, 2, \ldots, x\}$, in addition to its label $\psi(\epsilon)$.

G is an edge-labeled, directed graph. If we ignore direction, each directed graph is a member of a class of directed graphs all equivalent to their underlying undirected form created by symmetric closure. This is shown on the center left side of Fig. 1 for G. Alternatively, if we ignore labeling, each labeled graph is a member of a class of labeled graphs all equivalent to their underlying unlabeled form, as shown in the center right side of Fig. 1. Finally, the unlabeled, undirected form is shown on the bottom of Fig. 1.

We say that $H \subseteq G$ is a subgraph of G if every edge in H is also in G, so that $E_H \subseteq E_G$.[5] We restrict ourselves only to connected subgraphs $H \subseteq G$ of the connected graph G. A **motif** is then a collection of subgraphs which are equivalent by some criteria, and a k**-motif** is a motif all of whose subgraphs are of size $k \leq N$. In this work, we consider motifs which are equivalent by graph structure (directed or undirected), by labeling, and by both labeling and structure.

Fig. 2 enumerates the unlabeled (directed and undirected) motifs of size $k = 2, 3, 4$. For $k = 2, 3$, all directed motifs are enumerated; for $k = 4$, only those directed motifs present in G are shown. Each directed motif is identified by a motif number $m \in [[31]]$. Each motif maps to a collection of subgraphs $H \subseteq G$; Fig. 2 also shows the number $f(\mathcal{M})$ of those for both the directed and undirected (unlabeled) forms.

But there are structural relationships between the motifs, in that certain motifs of size k are contained within others of size $k + 1$, etc. The unlabeled, undirected case for our example is illustrated in Fig. 3. Each graph in the diagram represents a motif $\mathcal{M}$, in this case an unlabeled, undirected subgraph of the unlabeled, undirected form of G. Here we can now see more of the frequencies, ranging from just the edge count $N = 8$ for the single 1-motif to 1 for the single N-motif (the original graph).

The structure in Fig. 3 is a graded partially ordered set (poset), ordered by edge inclusion, ranked by k, and weighted by frequencies f. We recognize the unweighted form as a simplicial complex of subgraphs [2, 6, 7], restricted to the connected subgraphs. Simplicial complexes are familar as the structure formed by enumerating the k-dimensional hyper-faces of an N-dimensional polytope (multi-dimensional polygon) for $1 \in [[N]]$. The sub-graphs $H \in \mathcal{M}$ within each motif are thereby structurally (homotopically) equivalent.

Fig. 4 illustrates the unlabeled, directed case for $k \in [[4]]$. Note that each motif graph in Fig. 3 now expands to an equivalence class of directed motifs, as identified in the figure. It can be verified that the frequencies in the blocks add up to the frequencies for the motifs in Fig. 3, and indeed Fig. 3 is a sub-poset of Fig. 4.

Consider the motif identified as $\mathcal{M}^*$ at the top of Fig. 4 of size $k = 3$ and frequency $f(\mathcal{M}^*) = 3$. This motif alone is expanded to its full labeled, directed form in Fig. 5, along with its ancestors and descendants for $k \in [[4]]$ in the poset of labeled, directed motifs. As before, each motif in Fig. 4 is now expanded to its equivalence class.

3. MOTIF FREQUENCIES

[3]http://challenge.semanticweb.org

[4]Technically, graphs G used in this paper are also node-labeled, but always with unique labels serving only as identifiers. We also assume edges are singly-labeled. Extensions to the case where nodes are (properly) labeled, and $\psi \colon E \to 2^L$, so that G is an edge multigraph, can be considered.

[5]Note that this edge-based definition means that it could be that $V_H = V_G$ even if $E_H \subset E_G$.

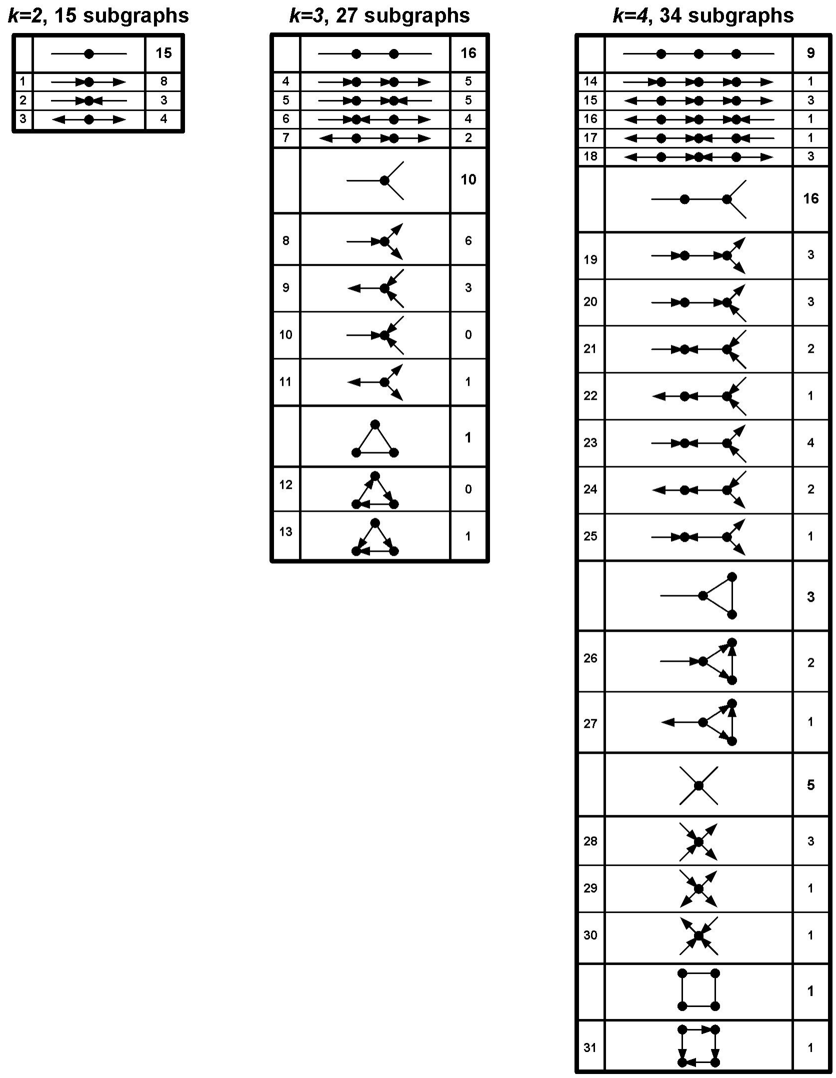

Figure 2: All motifs for $k = 2, 3, 4$, both the undirected and directed forms in its equivalence class. For each motif (undirected or directed), the right column shows its count, and for directed motifs the left column shows the motif # m. For $k = 4$, only those motifs are shown which are actually present in the graph in Fig. 1.

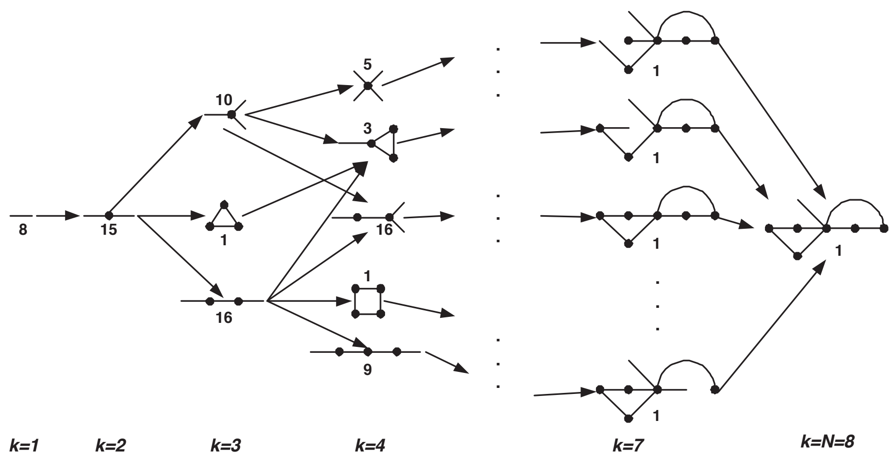

Figure 3: A portion of the unlabeled, undirected motif poset for our example.

Edge Label $l = \psi(\epsilon)$	Count	$p(l)$
f	4	0.500
g	3	0.375
h	1	0.125

Table 1: Edge label statistics for G.

Consider now the motifs of size $k = 1$ in Fig. 5. Of course, these are just the individual edge labels f, g, and h with frequencies $2, 2$, and 1 respectively. Space does not allow showing the full poset for the labeled, directed case, Fig. 5 is restricted to expand only the unlabeled, directed motif $\mathcal{M}^*$. Thus the frequencies for the motifs other than those of size $k = 3$ do not necessarily correspond to those of the classes shown in Fig. 4, since those also have contributions from its siblings in Fig. 4. So the frequencies for size $k = 1$ in Fig. 5 are actually those for the labels of edges which appear *within* the unlabeled, directed motif $\mathcal{M}^*$.

Removing this restriction, Table 1 shows the frequency distribution of the edge labels in the whole graph G, along with their *relative* frequencies $p\colon L \to [0,1]$, where for $l \in L, p(l) = f(l)/N$, so that $p(f) = .5, p(g) = .375, p(h) = .125$.

Any edge $\epsilon = \langle x, y\rangle \in E$ individually maps to the $k = 1$ unlabeled, directed motif $H = \langle \{x, y\}, \{\epsilon\}\rangle \subseteq G$. So just as each edge $\epsilon \in E$ has its label $\psi(\epsilon) \in L$, we seek to extend this concept to describe the **motif label** $\psi(H)$ of a whole subgraph $H \subseteq G$. For linear motifs, that is, motifs which are just paths, it is sufficient to use the vector of edge labels for $\psi(H)$. For example, for the two motifs of type $m = 1$ in Fig. 5, we have the vectors $\psi(\mathcal{M}) = \langle f, g\rangle$ and $\psi(\mathcal{M}) = \langle f, h\rangle$ respectively. In their undirected form, these would be sets $\psi(\mathcal{M}) = \{f, g\}$ and $\psi(\mathcal{M}) = \{f, h\}$.

But for non-linear motifs, $\psi(H)$ needs to be effectively the whole motif graph, indicating both the set of edge labels *and their connections*. Completion of this aspect awaits further work, but modulo these considerations, we can extend our notion of frequencies of edges $p(l) = p(\psi(\epsilon))$ to frequencies of motifs $p(\psi(H))$, or just $p(H)$ when clear from context. In particular, all of the directed and undirected motifs $H \subseteq G$ shown in Fig. 2 now break down into their labeled forms. Table 2 shows the frequency distribution $p(\psi(H))$ of the undirected motifs for $k = 2$, appropriately as sets of edge labels of size $k = 2$. In contrast, Table 3 shows the frequency distributions $p(\psi(H)|m)$ of the three directed motif patterns $m = 1, 2, 3$ for $k = 2$, appropriately as ordered pairs.

Undirected motif label $\psi(H)$	Count	$p(\psi(H))$
$\{f, f\}$	2	0.13
$\{f, g\}$	7	0.47
$\{f, h\}$	2	0.13
$\{g, g\}$	1	0.07
$\{g, h\}$	3	0.20
$\{h, h\}$	0	0.00

Table 2: Undirected motif label frequencies, $k = 2$.

Note the differences in the columns in Table 3, as only non-isomorphic pairs are listed. For motif $m = 1$, all $\eta^2 = 9$ combinations of edge labels are viable. But for motifs $m = 2, 3$, the patterns $\langle f, g\rangle$ and $\langle g, f\rangle$ are isomorphic, so that there are only $\eta^2 - \binom{\eta}{2} = 6$ possibilities.

4. INFORMATION GAIN OF MOTIFS

Now consider a labeled graph motif $H \subseteq G$, directed or undirected. We can count the frequency $p(H)$ as shown above. But we can also estimate how likely H is to occur at random given just the basic distribution $p(l)$ over labels. Let $\hat{p}(H)$ be this estimate of the expected frequency of $H \subseteq G$. Then for a measure of the **information gain** of the motif H, we use the standard the logarithmic form $-\log(p)$ to measure the information content of a probability $p \neq 0$. Noting that $p \geq q \to -\log(p) \leq -\log(q)$, we posit

$$I(H) := \log(p(H)) - \log(\hat{p}(H)) = \log \frac{p(H)}{\hat{p}(H)},$$

for both $p(H), \hat{p}(H) \neq 0$. $I(H)$ measures the amount to which $p(H)$ occurs above its expectation $\hat{p}(H)$, so that then $I(H) > 0$, and $I(H) < 0$ if it occurs less than expected.

To estimate $\hat{p}(H)$ from $p(l)$ for a directed linear motif $H \subseteq G$, it is sufficient to let $\hat{p}(H) = \prod_{\epsilon \in H} p(\psi(\epsilon))$, where we iterate over each of the edges ϵ which compose H. The results for the first directed linear motif pattern $m = 1$ are shown in Table 4. Note that Table 4 is restricted to only those motifs which occur in the graph. These are all we're measuring, and this guarantees that $p(H), \hat{p}(H) > 0$.

For undirected (linear) motifs, calculating $\hat{p}(H)$ is more complicated. For $k = 2$ only, let $H = \{l_1, l_2\}$ be the motif label, consisting of its two distinct edge labels $l_1 = \psi(\epsilon_1), l_2 =$

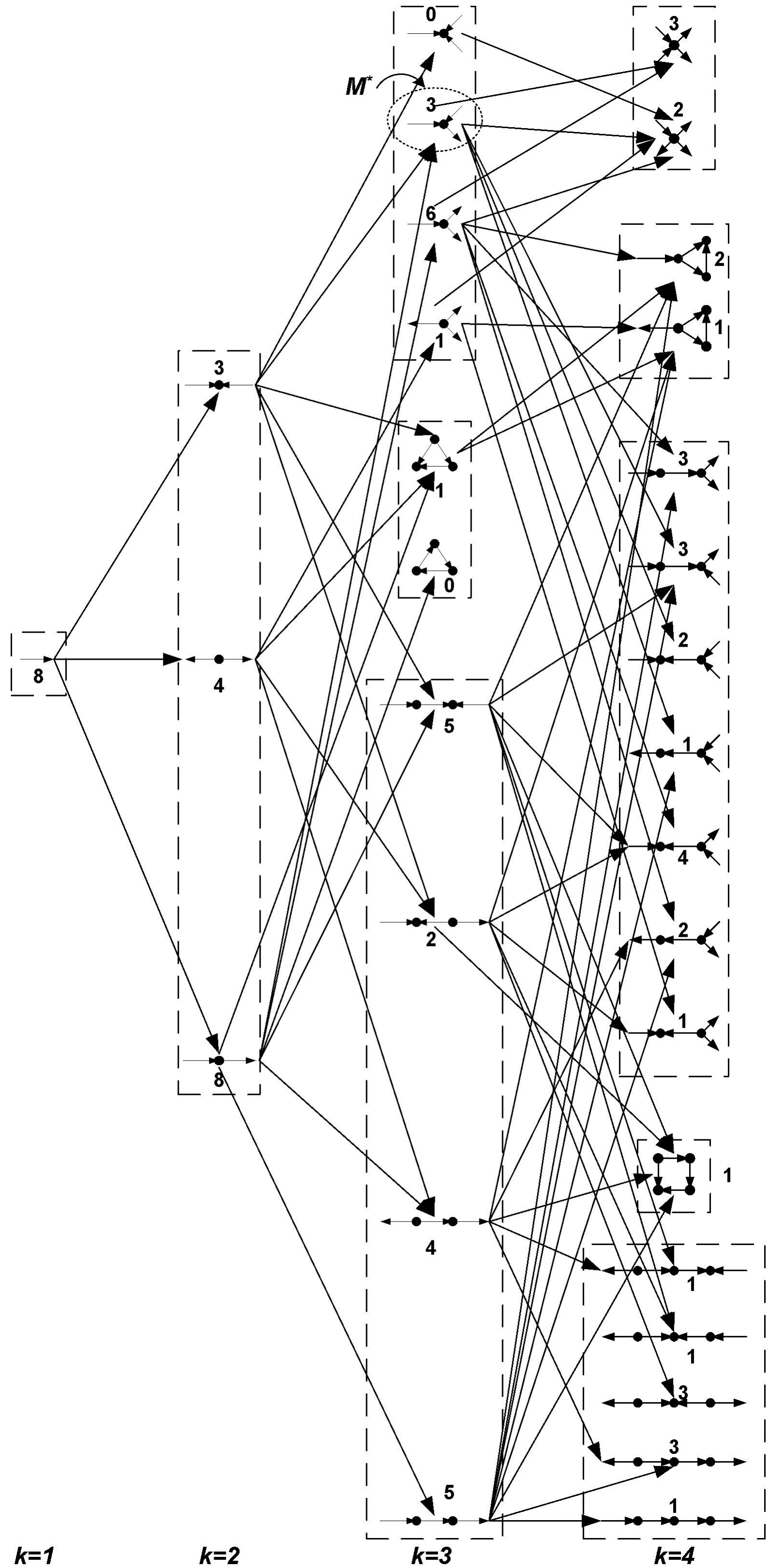

Figure 4: The unlabeled, directed motif poset set for our example for $k \in [[4]]$.

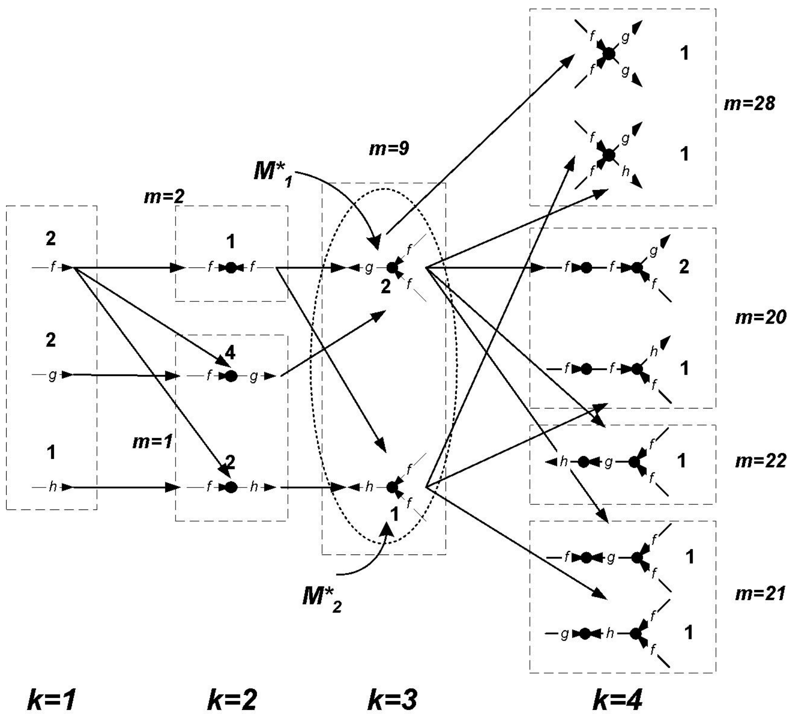

Figure 5: A portion of the labeled, directed motif poset implied by the specific unlabeled, directed motif $\mathcal{M}^*$.

Directed motif label	$m = 1$		$m = 2$		$m = 3$	
$\psi(H)$	Count	$p(\psi(H)\|m)$	Count	$p(\psi(H)\|m)$	Count	$p(\psi(H)\|m)$
$\langle f, f\rangle$	1	0.125	1	0.333	0	0.000
$\langle f, g\rangle$	4	0.500	1	0.333	1	0.250
$\langle f, h\rangle$	2	0.250	0	0.000	0	0.000
$\langle g, f\rangle$	1	0.125				
$\langle g, g\rangle$	0	0.000	0	0.000	1	0.250
$\langle g, h\rangle$	0	0.000	1	0.333	2	0.500
$\langle h, f\rangle$	0	0.000				
$\langle h, g\rangle$	0	0.000				
$\langle h, h\rangle$	0	0.000	0	0.000	0	0.000

Table 3: Directed motif label frequencies for $k = 2$, motifs $m = 1, 2, 3$.

Directed linear motif label	Count	$p(\psi(H))$	$\hat{p}(H)$	$I(H)$
$\langle f, f\rangle$	1	0.125	0.250	-0.301
$\langle f, g\rangle$	4	0.500	0.188	0.426
$\langle f, h\rangle$	2	0.250	0.063	0.602
$\langle g, f\rangle$	1	0.125	0.188	-0.176

Table 4: Gains for directed motifs, $k = 2, m = 1$.

Undirected linear motif label	Count	$p(H)$	$\hat{p}(H)$	$I(H)$
$\{f, f\}$	2	0.133	0.250	-0.273
$\{f, g\}$	7	0.467	0.375	0.095
$\{f, h\}$	2	0.133	0.125	0.028
$\{g, g\}$	1	0.067	0.141	-0.323
$\{g, h\}$	3	0.200	0.094	0.329

Table 5: Gains for undirected motifs, $k = 2$.

$\psi(\epsilon_2)$. Then we have:

$$\hat{p}(H) = \begin{cases} (p(l))^2, & l = l_1 = l_2 \\ 2 \cdot p(l_1) \cdot p(l_2), & l_1 \neq l_2 \end{cases} \quad (1)$$

Table 5 shows the results for undirected motifs for $k = 2$. The highest information gain is for the motif $\{g, h\}$, since it would be expected to occur 9.4% of the time, when in fact it occurs 20.0% of the time. Similarly, $\{f, f\}$ is underrepresented at 13.3% compared to 25.0%. For $k > 2$, (1) becomes substantially more complicated

Consider our target motifs $\mathcal{M}_1^*, \mathcal{M}_2^*$ again from Fig. 5, occurring with empirical frequencies of $p(H(\mathcal{M}_1^*)) = 2/3$ and $p(H(\mathcal{M}_2^*)) = 1/3$ respectively within their classes. Each is of motif type $m = 9$, and is composed of types $m = 1, 2$. For example, $\mathcal{M}_1^*$ is the union of $m = 2$ with label $\psi = \langle f, f\rangle$ and $m = 1$ with label $\psi = \langle f, g\rangle$. Thus we can posit the estimated expected frequencies $\hat{p}(H(\mathcal{M}_1^*)), \hat{p}(H(\mathcal{M}_2^*))$ as the product of those constituent frequencies, which are shown on Table 3 as $\hat{p}(H(\mathcal{M}_1^*)) = 1/2 \cdot 1/3 = 1/6$ and $\hat{p}(H(\mathcal{M}_2^*)) = 1/12$. The corresponding information gains are then $I(\mathcal{M}_1^*) = \log 6 = 0.78, I(\mathcal{M}_2^*) = \log 4 = 0.60$.

5. BILLION TRIPLE CHALLENGE TRIGRAMS

We now show information gains for some linear motifs from BTC10 [8, 9], an RDF graph with 1.4B unique $\langle s, p, o\rangle$ triples. Fig. 6 shows the top 16 of the 95.2K predicates, comprising 35% of all 1.4B link instances, as shown by the cumulative percentage line. Fig. 7 shows the EO for the top 30 edge labels in BTC10. For example we have about 70M triples with the predicate `foaf:knows` connecting subject and object of class `foaf:Person`, the highest count.

Table 6 shows the distribution of the top 20 bigrams of the 1.3M consecutive link type pairs, comprising 53.0% of all 17.0B consecutive link pairs present; and Table 7 shows the distribution of the top 20 trigrams of the 72.7M consecutive predicate triples, comprising 7.54% of all 1.04T link triples.

Note that there is a subtle formal difference between the EO in Fig. 7 and our example in Fig. 1. Both have node labels which are unique identifiers, in the EO these are the classes of resources. Edge labels in the EO are predicates. But the EO is additionally weighted by the counts of the edges, basically aggregating multiple edges of the form $x \xrightarrow{f} y$ into one edge with the weight being its count. This would be equivalent to our Fig. 1 being a multi-graph. In any event, there is no difference once counts are made strictly on link types (edge labels) in both structures, these are just two different mechanisms to add up counts of link types.

Low-frequency predicates are prominent in both the bigrams and trigrams. For example, consider the most frequent bigram $\langle$`dgtwc:isPartOf, dgtwc:partial_data`$\rangle$, with a frequency of 17.1%. The constituent predicates have frequencies of 0.0038% and 0.027% respectively, far below the top 16 shown in Fig. 6. If these were independent, the expected joint frequency would be minuscule. For information gains, we have $I = 7.22$. This pattern of a vast inflation of expected probability is a general phenomenon, indicating the powerful role that these small sequence motifs play in the semantics of BTC10.

Table 8 shows information gains for the top 7 link type trigrams. Note that the third and fourth rows are structurally isomorphic (trigram motifs $\langle f, g, f\rangle$ and $\langle g, f, g\rangle$ have the same decomposition into bigram motifs $\langle f, g\rangle$ and $\langle g, f\rangle$), so their counts are combined into the third row of Table 8.

These initial results are insufficient to draw conclusions, but we can see that there is a significant variation in I for the different trigrams, and a lack of obvious dependence between $I(H)$ values and base motif frequency $p(H)$. This is initial justification in the value of $I(H)$ to indicate additional information not present in the base frequencies.

6. FURTHER WORK

We have shown preliminary results on the use of information theoretical measures to assess the significance of edge-labeled motifs in semantic graph databases. A number of developments await immediate progress beyond this first workshop paper:

- We recognize that our mathematical objects have been explored in combinatorics and algebraic topology, and we seek results from simplicial complexes which we can bring to bear. We can see I as an objective function in a combinatorial search problem over these posets. In particular, we are interested in exploiting constraint relationships which exist on the frequencies f, p of particular motifs in terms of the frequencies of their children and parents in the posets.

- We also need to extend our understanding of the expressions for the expected frequencies $\hat{p}(H)$ for nonlinear motifs. Simply taking the product of constituents does not completely reflect the structural overlap.

- Additional interaction between our EO approach and this measurement method is also in order. In particular, in real RDF graphs nodes can have multiple types. Possible approaches then include making our input graph node-weighted, or multi-node-labeled. But there could be edges between *different* nodes of the *same* type participating in *different* motifs in the EO. We may seek to expand the EO to accommodate this, thus counting motifs at the instance level.

- Finally, we are straining SUBDUE by using it for new purposes. Additional software development will be very useful, and there is active work underway by our team to scale SUBDUE to giga-scale levels.

7. ACKNOWLEDGEMENTS

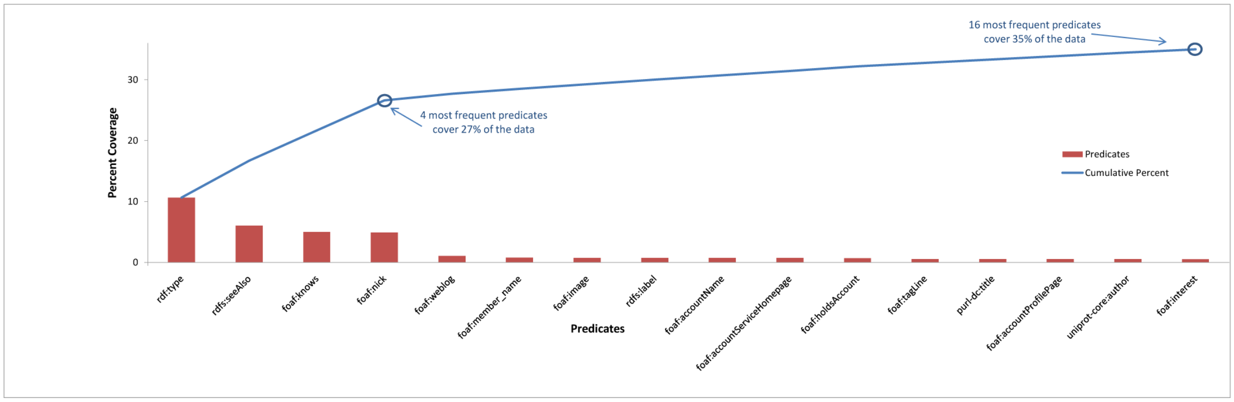

Figure 6: Top 16 predicates in BTC10.

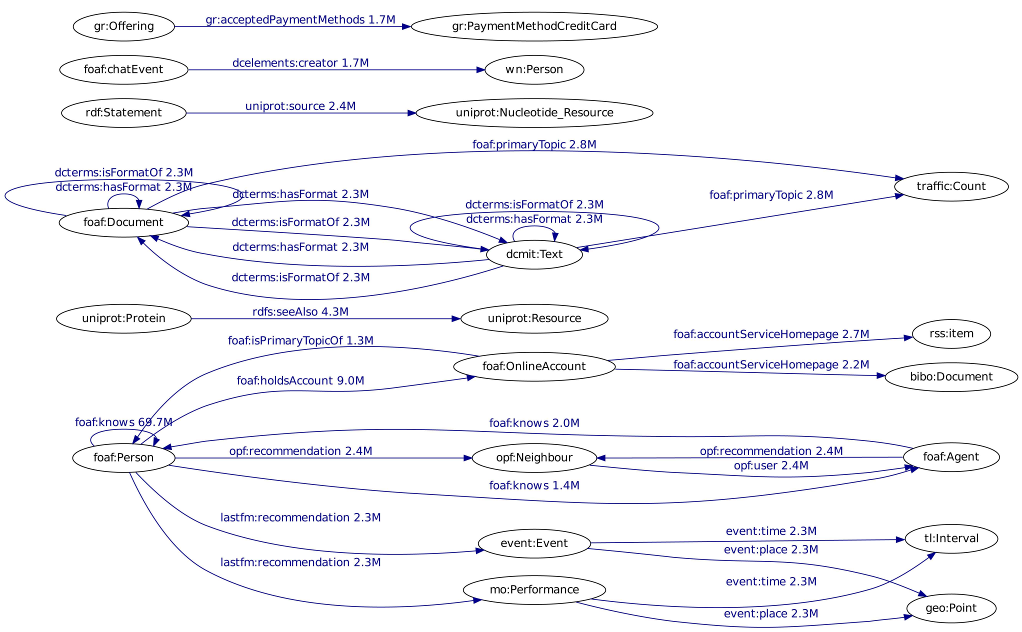

Figure 7: The extant Ontology for the Top 30 Link-node-types in BTC10.

l_1	l_2	Count (M)	%
dgtwc:isPartOf	dgtwc:partial_data	2912.13	35.8%
foaf:interest	purl:title	2036.23	25.0%
gs:isUnknownAboutIn	gs:hasUnknownExpectationOf	516.19	6.3%
gs:isExpectedIn	gs:hasExpectationOf	142.62	1.8%
gs:isUnknownAboutIn	gs:hasLowExpectationOf	139.13	1.7%
gs:isUnexpectedIn	gs:hasUnknownExpectationOf	139.13	1.7%
gs:isUnknownAboutIn	gs:hasExpectationOf	132.04	1.6%
gs:isExpectedIn	gs:hasUnknownExpectationOf	132.04	1.6%
gs:isUnexpectedIn	gs:hasLowExpectationOf	124.14	1.5%
sioc:follows	sioc:follows	116.87	1.4%
gs:isUnexpectedIn	gs:hasExpectationOf	84.12	1.0%
gs:isExpectedIn	gs:hasLowExpectationOf	84.12	1.0%
foaf:knows	foaf:knows	81.69	1.0%
foaf:primaryTopic	foaf:maker	77.68	1.0%
foaf:knows	foaf:nick	68.93	0.8%
fao:hasScope	fao:isScopeOf	60.16	0.7%
fao:hasType	fao:isTypeOf	52.24	0.6%
foaf:accountServiceHomepage	purl:title	41.69	0.5%
foaf:knows	foaf:holdsAccount	39.26	0.5%
foaf:based_near	gs:hasUnknownExpectationOf	36.16	0.4%

Table 6: Top 20 link type bigrams in BTC10 (millions).

l_1	l_2	l_3	Count (B)	%
sioc:follows	sioc:follows	sioc:follows	10.85	10.6%
foaf:knows	foaf:knows	foaf:knows	2.19	2.1%
gs:hasUnknownExpectationOf	gs:isUnknownAboutIn	gs:hasUnknownExpectationOf	2.15	2.1%
gs:isUnknownAboutIn	gs:hasUnknownExpectationOf	gs:isUnknownAboutIn	2.15	2.1%
gs:isUnknownAboutIn	gs:hasUnknownExpectationOf	foaf:isPrimaryTopicOf	1.98	1.9%
gs:isUnknownAboutIn	gs:hasUnknownExpectationOf	skos:closeMatch	1.37	1.3%
rdf:predicate	http://sw.nokia.com/VOC-1/partOf	http://sw.nokia.com/VOC-1/term	1.32	1.3%
foaf:primaryTopic	gs:isUnknownAboutIn	gs:hasUnknownExpectationOf	1.08	1.1%
skos:closeMatch	gs:isUnknownAboutIn	gs:hasUnknownExpectationOf	0.78	0.8%
gs:hasUnknownExpectationOf	gs:isUnknownAboutIn	gs:hasExpectationOf	0.64	0.6%
gs:isExpectedIn	gs:hasUnknownExpectationOf	gs:isUnknownAboutIn	0.64	0.6%
gs:hasUnknownExpectationOf	gs:isUnknownAboutIn	gs:hasLowExpectationOf	0.64	0.6%
gs:isUnexpectedIn	gs:hasUnknownExpectationOf	gs:isUnknownAboutIn	0.64	0.6%
gs:isExpectedIn	gs:hasExpectationOf	foaf:isPrimaryTopicOf	0.62	0.6%
gs:isUnknownAboutIn	gs:hasLowExpectationOf	foaf:isPrimaryTopicOf	0.62	0.6%
gs:isUnknownAboutIn	gs:hasExpectationOf	foaf:isPrimaryTopicOf	0.61	0.6%
gs:isUnknownAboutIn	gs:hasLowExpectationOf	gs:isUnexpectedIn	0.56	0.5%
gs:hasLowExpectationOf	gs:isUnexpectedIn	gs:hasUnknownExpectationOf	0.56	0.5%
gs:isUnexpectedIn	gs:hasLowExpectationOf	foaf:isPrimaryTopicOf	0.53	0.5%
gs:isUnknownAboutIn	gs:hasUnknownExpectationOf	gs:isExpectedIn	0.52	0.5%

Table 7: Top 20 link type trigrams in BTC10 (billions).

l_1	l_2 l_3	$p(H)$	$p(\langle l_1, l_2\rangle)$	$p(\langle l_2, l_3\rangle)$	$\hat{p}(H)$	$I(H)$
sioc:follows	sioc:follows sioc:follows	10.60%	1.44%	1.44%	0.021%	2.71
foaf:knows	foaf:knows foaf:knows	2.13%	1.00%	1.00%	0.010%	2.33
gs:hasUnknownExpectationOf	gs:isUnknownAboutIn gs:hasUnknownExpectationOf	2.09%	0.00091%	6.35%	0.00012%	4.26
gs:isUnknownAboutIn	gs:hasUnknownExpectationOf foaf:isPrimaryTopicOf	1.93%	6.35%	0.00080%	0.000051%	4.58
gs:isUnknownAboutIn	gs:hasUnknownExpectationOf skos:closeMatch	1.33%	6.35%	0.00055%	0.000035%	4.58
rdf:predicate	http://sw.nokia.com/VOC-1/partOf http://sw.nokia.com/VOC-1/term	1.28%	0.0216%	0.00367%	0.00000079%	6.21
foaf:primaryTopic	gs:isUnknownAboutIn gs:hasUnknownExpectationOf	1.05%	0.00044%	6.35%	0.000028%	4.57

Table 8: Information gains for top 6 link type trigrams in BCT10, $H = \langle l_1, l_2, l_3\rangle$.

Thanks to Emilie Hogan for a critical review. This work was funded in part by the Center for Adaptive Supercomputing Software – Multithreaded Architectures (CASS-MT) at the Dept. of Energy's Pacific Northwest National Laboratory. Pacific Northwest National Laboratory is operated by Battelle Memorial Institute under Contract DE-ACO6-76RL01830.

8. REFERENCES

[1] al-Saffar, Sinan; Joslyn, Cliff A; and Chappell, Alan: (2011) "Extant Ontological Scaling and Descriptive Semantics for Semantic Structure Discovery in Large Graph Datasets", *IEEE/WIC/ACM Int. Conf. on Web Intelligence*

[2] Björner, Anders and Welker, Volkmar: (1999) "Complexes of Directed Graphs", *SIAM J. Discrete Mathematics*, v. **12**, pp. 413-424

[3] Cook, Diane J and Holder, Lawrence B: (2000) "Graph-Based Data Mining", *IEEE Intelligent Systems*, v. **15**:2, pp. 32-41

[4] Cook, Diane J and Holder, Lawrence B, eds.: (2007) *Mining Graph Data*, Wiley

[5] Itzkovitz, S; Milo, R; Kashtan, N; Ziv, G; Alon, U: (2003) "Subgraphs in Random Networks", *Physical Review E*, v. **68**

[6] Jonsson, Jakob: (2007) *Simplicial Complexes of Subgraphs of a Graph on At Most Six Vertices*, `http://www.math.kth.se/~jakobj/doc/manuscripts/6.pdf`

[7] Jonsson, Jakob: (2008) *Simplicial Complexes of Graphs*, Springer-Verlag, Berlin

[8] Joslyn, Cliff; Adolf, Bob; al-Saffar, Sinan; Feo, J; and David Haglin: (2010) "High Performance Semantic Factoring of Giga-Scale Semantic Graph Databases", in: *Semantic Web Challenge 2010, Int. Semantic Web Conf.*, `http://www.cs.vu.nl/~pmika/swc/submissions/swc2010_submission_15.pdf`

[9] Joslyn, Cliff; Adolf, Bob; al-Saffar, Sinan; Feo, J; and David Haglin: (2011) "High Performance Descriptive Semantic Analysis of Semantic Graph Databases", in: *1st Wshop. on High-Performance Computing for the Semantic Web (HPCSW 2011)*

[10] Schmidt, Michael; Hornung, Thomas; Küchlin, Norbert; Lausen, G; Christoph Pinkel: (2008) "An Experimental Comparison of RDF Data Management Approaches in a SPARQL Benchmark Scenario", in: *Proc. 7th Int. Conf. Semantic Web (ISWC 08)*, pp. 82-97, `doi>10.1007/978-3-540-88564-1_6`

[11] Zou, Ruoyu and Holder, Lawrence B: (2010) "Frequent Subgraph Mining on a Single Large Graph Using Sampling Techniques", in: *Proc. 8th Wshop. Mining and Learning with Graphs (MLG 10)*, `doi>10.1145/1830252.1830274`