ABSTRACT

Given a semantic graph data set, perhaps one lacking in an explicit ontology, we try to 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 datasets are sets of triples \((s, p, o)\), 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 \(C(s), C(o)\), 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] 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.

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

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires specific permission and/or a fee.

MDS 2011 Prepared for the Mining Data Semantics Workshop.
Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ... $10.00.

1http://ailab.wsu.edu/subdue
2. ORDERED SETS OF MOTIFS IN LABELED, DIRECTED GRAPHS

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

\( G \) 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 \( E_H \subseteq E_G \). We restrict ourselves only to connected 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 \( f(M) \) 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 \( M \), 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 familiar 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 \( H \subseteq M \) 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 \( M^* \) at the top of Fig. 4 of size \( k = 3 \) and frequency \( f(M^*) = 3 \). 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
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.
consider the frequency distribution of the edge labels in the whole graph $G$, along with their relative frequencies $p: L \rightarrow [0,1]$, where for $l \in L$, $p(l) = f(l)/N$, so that $p(f) = .5, p(g) = .375, p(h) = .125$. Any edge $e = (x,y) \in E$ individually maps to the $k = 1$ unlabeled, directed motif $H = \langle\{x,y\}, \{e\}\rangle \subseteq G$. So just as each edge $e \in E$ has its label $\psi(e) \in L$, 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 $\psi(M) = (f,g)$ and $\psi(M) = (f,h)$ respectively. In their undirected form, these would be sets $\psi(M) = \{f,g\}$ and $\psi(M) = \{f,h\}$.

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 $p(l) = p(\psi(e))$ 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.

Note the differences in the columns of 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 $(f,g)$ and $(g,f)$ are isomorphic, so that there are only $\eta^2 - \binom{2}{2} = 6$ 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 $\bar{p}(H)$ 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 logarithmic form $-\log(p)$ to measure the information content of a probability $p \neq 0$. Noting that $p \geq q \implies -\log(p) \leq -\log(q)$, we posit

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

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

To estimate $\bar{p}(H)$ from $p(l)$ for a directed linear motif $H \subseteq G$, it is sufficient to let $\bar{p}(H) = \prod_{e \in H} p(\psi(e))$, where we iterate over each of the edges $e$ 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 $p(H), \bar{p}(H) > 0$.

For undirected (linear) motifs, calculating $\bar{p}(H)$ is more complicated. For $k = 2$ only, let $H = \{l_1, l_2\}$ be the motif label, consisting of its two distinct edge labels $l_1 = \psi(e_1), l_2 = \psi(e_2)$. 

<table>
<thead>
<tr>
<th>Edge Label</th>
<th>Count</th>
<th>$p(l)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f$</td>
<td>4</td>
<td>0.500</td>
</tr>
<tr>
<td>$g$</td>
<td>3</td>
<td>0.375</td>
</tr>
<tr>
<td>$h$</td>
<td>1</td>
<td>0.125</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Undirected motif label</th>
<th>Count</th>
<th>$p(\psi(H))$</th>
</tr>
</thead>
<tbody>
<tr>
<td>${f,f}$</td>
<td>2</td>
<td>0.13</td>
</tr>
<tr>
<td>${f,g}$</td>
<td>7</td>
<td>0.47</td>
</tr>
<tr>
<td>${f,h}$</td>
<td>2</td>
<td>0.13</td>
</tr>
<tr>
<td>${g,g}$</td>
<td>1</td>
<td>0.07</td>
</tr>
<tr>
<td>${g,h}$</td>
<td>3</td>
<td>0.20</td>
</tr>
<tr>
<td>${h,h}$</td>
<td>0</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 1: Edge label statistics for $G$. 

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

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

<table>
<thead>
<tr>
<th>Directed motif label</th>
<th>$m = 1$</th>
<th>$m = 2$</th>
<th>$m = 3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\psi(H)$</td>
<td>Count</td>
<td>$p(\psi(H)</td>
<td>m)$</td>
</tr>
<tr>
<td>$(f, f)$</td>
<td>1</td>
<td>0.125</td>
<td>1</td>
</tr>
<tr>
<td>$(f, g)$</td>
<td>4</td>
<td>0.500</td>
<td>1</td>
</tr>
<tr>
<td>$(f, h)$</td>
<td>2</td>
<td>0.250</td>
<td>0</td>
</tr>
<tr>
<td>$(g, f)$</td>
<td>1</td>
<td>0.125</td>
<td>0</td>
</tr>
<tr>
<td>$(g, g)$</td>
<td>0</td>
<td>0.000</td>
<td>0</td>
</tr>
<tr>
<td>$(g, h)$</td>
<td>0</td>
<td>0.000</td>
<td>1</td>
</tr>
<tr>
<td>$(h, f)$</td>
<td>0</td>
<td>0.000</td>
<td>0</td>
</tr>
<tr>
<td>$(h, g)$</td>
<td>0</td>
<td>0.000</td>
<td>0</td>
</tr>
<tr>
<td>$(h, h)$</td>
<td>0</td>
<td>0.000</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3: Directed motif label frequencies for $k = 2$, motifs $m = 1, 2, 3$. 
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 bi-
grams and trigrams. For example, consider the most fre-
quent bigram \( \langle \text{dgtwc:isPartOf}, \text{dgtwc:partial_data} \rangle \), with a
frequency of 17.1%. The constituent predicates have fre-
quencies 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 tri-
grams. Note that the third and fourth rows are structurally
isomorphic (trigram motifs \( \langle f, g, f \rangle \) and \( \langle g, f, f \rangle \) 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 be-
tween \( I(H) \) 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 infor-
mation theoretical measures to assess the significance of
edge-labeled motifs in semantic graph databases. A num-
ber 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 par-
ticular motifs in terms of the frequencies of their chil-
dren and parents in the posets.

- We also need to extend our understanding of the ex-
pressions for the expected frequencies \( \hat{p}(H) \) for non-
linear 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 par-
ticular, in real RDF graphs nodes can have multiple
types. Possible approaches then include making our in-
put 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 strainung 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
Table 6: Top 20 link type bigrams in BTC10 (millions).
Table 7: Top 20 link type trigrams in BTC10 (billions).

<table>
<thead>
<tr>
<th>l₁</th>
<th>l₂</th>
<th>l₃</th>
<th>p(l₁)</th>
<th>p(l₂)</th>
<th>p(l₁, l₂)</th>
<th>p(l₁, l₂, l₃)</th>
<th>Count (B)</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>foaf:follows</td>
<td>foaf:knows</td>
<td>foaf:knows</td>
<td>0.000028%</td>
<td>0.000044%</td>
<td>6.35%</td>
<td>0.0000079%</td>
<td>6.21</td>
<td></td>
</tr>
<tr>
<td>ge:hasUnknownExpectationOf</td>
<td>ge:hasUnknownExpectationOf</td>
<td>ge:hasUnknownExpectationOf</td>
<td>0.00091%</td>
<td>0.00091%</td>
<td>6.35%</td>
<td>0.000098%</td>
<td>4.58</td>
<td></td>
</tr>
<tr>
<td>ge:hasUnknownExpectationOf</td>
<td>ge:hasUnknownExpectationOf</td>
<td>ge:hasUnknownExpectationOf</td>
<td>0.00012%</td>
<td>0.00012%</td>
<td>6.35%</td>
<td>0.000005%</td>
<td>4.58</td>
<td></td>
</tr>
<tr>
<td>ge:hasUnknownExpectationOf</td>
<td>ge:hasUnknownExpectationOf</td>
<td>ge:hasUnknownExpectationOf</td>
<td>0.000051%</td>
<td>0.000051%</td>
<td>6.35%</td>
<td>0.000035%</td>
<td>4.58</td>
<td></td>
</tr>
<tr>
<td>ge:hasUnknownExpectationOf</td>
<td>ge:hasUnknownExpectationOf</td>
<td>ge:hasUnknownExpectationOf</td>
<td>0.000035%</td>
<td>0.000035%</td>
<td>6.35%</td>
<td>0.000005%</td>
<td>4.58</td>
<td></td>
</tr>
<tr>
<td>ge:hasUnknownExpectationOf</td>
<td>ge:hasUnknownExpectationOf</td>
<td>ge:hasUnknownExpectationOf</td>
<td>0.0000079%</td>
<td>0.0000079%</td>
<td>6.35%</td>
<td>0.000002%</td>
<td>4.57</td>
<td></td>
</tr>
</tbody>
</table>

Table 8: Information gains for top 6 link type trigrams in BCT10, $H = \{l₁, l₂, l₃\}$.

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-AC06-76RL01830.

8. REFERENCES

[1] al-Saﬃr, Sinan; Joslyn, Cliff; 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


