1 Attribute distributions on undirected graphs

Suppose we have an undirected graph $G = (V, E)$ a set A of attributes, and an attribute assignment function $\phi : V \to A$ mapping vertices to their attributes. For simplicity, we may assume that A contains a "missing" attribute token to cover the case when the attribute of a vertex is unknown or missing. We wish to compute a similarity score between pairs of attributes that measures their propensity to occur together in edges in the graph. In other words, we seek a similarity function $s : A \times A \to \mathbb{R}$ for which greater values of $s(a, b)$ indicate greater likelihood of co-occurrence of a and b. Furthermore, s should be symmetric, due to the undirected nature of the graph.

Following [1], we use *pointwise mutual information* (PMI) as a similarity score. PMI is defined as:

$$\mathrm{PMI}(a, b) = \log \frac{P(a, b)}{P(a)P(b)}.$$

This measures the relative co-occurrence likelihood of a and b with respect to the assumption of independence. If a and b occur independently of each other, then the PMI will be 0. The PMI will be positive (negative) if a and b occur more (less) frequently together than if they had been chosen independently.

$P(a, b)$ is defined as the probability that an edge in E will have attributes $\{a, b\}$. In the graph context, it is tempting to think of $P(a)$ as the probability that a vertex in V has attribute a, but this is incorrect. $P(a)$ is in fact the marginal probability that a occurs as one of the attributes of an edge $\{u, v\} \in E$:

$$P(a) = P\left(\bigcup_{b \in A} \{\phi(u), \phi(v)\} = \{a, b\}\right) = \sum_{b \in A} P(a, b).$$

2 Normalization of PMI

Assuming that $P(a) > 0$ for all $a \in A$, the range of values for $\mathrm{PMI}(a, b)$ is

$$[-\infty, \min(-\log P(a), -\log P(b))].$$

This is because PMI can also be written as $\log \frac{P(a \mid b)}{P(a)}$ or $\log \frac{P(b \mid a)}{P(b)}$, which are respectively maximized when $P(a \mid b) = 1$ or $P(b \mid a) = 1$, yielding $\min(-\log P(a), -\log P(b))$ as an upper bound. In the event that a and b only occur with each other, $P(a, b) = P(a) = P(b)$, and the upper bound reduces to $-\log P(a, b)$. Noting that this upper bound varies with a and b, we propose two alternative normalizations of PMI in the following table.

Similarity	**Range**	**Independence value**
$\mathrm{PMI}(a, b) = \log \frac{P(a,b)}{P(a)P(b)}$	$(-\infty, -\log P(a, b)]$	0
$\mathrm{NPMI1}(a, b) = \frac{\log P(a,b) - \mathrm{PMI}(a,b)}{2 \log P(a,b)} = \frac{\log(P(a)P(b))}{2 \log P(a,b)}$	$[0, 1]$	$\frac{1}{2}$
$\mathrm{NPMI2}(a, b) = -\frac{\log(1 - \mathrm{NPMI1}(a,b))}{\log 2}$	$[0, \infty)$	1

Depending on the application, using one of these may be preferable to using the basic PMI. In particular, NPMI1 is nonnegative and bounded, both of which may be desirable properties. Furthermore, each of these similarity scores can be readily converted to dissimilarity scores using appropriate shifts and reflections.

3 Empirical PMI

The formula for PMI treats the attributes on the graph vertices as random variables, but we wish to estimate the PMI from an actual graph whose attributes are known. To do so, we use the sample frequencies of attribute pairs in the edge set. Let

$$f_{a,b} = |\{\{u,v\} \in E : \{a,b\} = \{\phi(u),\phi(v)\}\}|, \quad f_a = |\{\{u,v\} \in E : a \in \{\phi(u),\phi(v)\}\}|.$$

Then we estimate $P(a,b)$ by $\hat{P}(a,b) = \frac{f_{a,b}}{|E|}$, and $P(a)$ by $\hat{P}(a) = \frac{f_a}{\sum_b f_b}$. We can then use $\hat{P}$ in place of P to compute the estimated PMI. Note that by definition, $\sum_{a,b} f_{a,b} = |E|$, meaning that $\hat{P}$ is a properly normalized probability.

As is often the case with real data, G will be sparse, and A may be a large set, many of whose elements may only be represented by just a few edges in E. In such cases, $f_{a,b} = 0$ for a vast majority of a, b pairs, but assigning these pairs an empirical probability of 0 will drastically underrepresent them. To give them positive probability, we smooth the counts by adding a constant $\delta > 0$ to each $f_{a,b}$. This gives the following smoothed versions of the empirical probabilities:

$$\hat{P}(a,b) = \frac{f_{a,b} + \delta}{|E| + \delta\binom{|A|+1}{2}}, \quad \hat{P}(a) = \frac{f_a + \delta|A|}{\sum_b f_b + \delta|A|^2}.$$

We see that $\hat{P}(a,b)$ has the appropriate normalization because the set of attribute pairs is of size $\binom{|A|+1}{2}$, since there are $\binom{|A|}{2}$ pairs of distinct attributes and $|A|$ pairs of identical attributes. Having $\delta > 0$ guarantees that all probabilities are positive, and increasing δ distorts the distribution closer to the uniform distribution on attribute pairs.

4 Sparse representation of Empirical PMI

Suppose we want to represent the empirical PMI matrix in a sparse way. Without smoothing, we could compute the empirical PMI for only the attribute pairs $\{a,b\}$ for which $f_{a,b} > 0$, so that the required time and memory is only $O(|E|)$ instead of $O(|A|^2)$. In practice, it may be best to use NPMI1 or NPMI2 so that all unrepresented attribute pairs have similarity 0 instead of $-\infty$.

If we want to use the δ-smoothing, things get a little more complicated. The empirical PMI matrix is not just a sum of a sparse matrix and a constant matrix (as would be the case for the counts alone) due to the $\hat{P}(a)\hat{P}(b)$ in the denominator. However, we can represent the matrix in the following way.

Let F be the matrix $F_{i,j} = \log(f_{a_i,a_j} + \delta) - \log(\delta)$, where $a_1, a_2, \ldots, a_{|A|}$ is some canonical ordering of the attributes. Note that $F_{i,j} = 0$ whenever $f_{a_i,a_j} = 0$, meaning we can represent F sparsely

when the pairwise counts are sparse. Now let $\mathbf{v}$ be the vector for which $v_i = \log(f_{a_i} + \delta|A|)$, and let $\Delta = \log(\delta) + \log(|E| + \delta\binom{|A|+1}{2})$. Denoting by M the matrix of empirical PMIs, we have

$$M = F + \Delta \mathbf{1}\mathbf{1}^T - \mathbf{v}\mathbf{1}^T - \mathbf{1}\mathbf{v}^T,$$

where $\mathbf{1}$ is the vector of all 1's. So M can be represented as a sum of a sparse matrix and three rank-one matrices. The memory requirement is $O(|E| + |A|)$, and the sparse representation will enable fast computation of various embedding/clustering algorithms.

The sparse sum representation of the empirical PMI is licensed by the linearity of PMI with respect to the log-probabilities. Such linearity does not exist for the normalized versions of PMI, so a different kind of smoothing would be required in which normalized PMIs are computed for all nonzero-frequency attribute pairs, and then a small constant is added to all the values *after* normalization.

5 Attribute-induced Kernels

At this point we have a sparse representation of a similarity kernel on A. This may be the matrix of empirical joint probabilities, or alternatively one of the empirical PMI variants described above. To perform inference on G, e.g. by a weighted graph embedding, we may need a similarity kernel on V rather than A.

Suppose $n \geq m$, $V = \{1, 2, \ldots, n\}$, $A = \{1, 2, \ldots, m\}$, and $\phi : V \to A$ maps vertices to attributes. Then we define an **collapse operator** S as the $m \times n$ matrix where

$$S_{ij} = I\{\phi(j) = i\}.$$

Let K be an $m \times m$ (symmetric) similarity kernel on A. Then we say $K' = S^T K S$ is the **attribute-induced kernel** on V. This kernel has the property

$$\begin{aligned}
K'_{ij} &= \sum_{a=1}^{m} S_{ai} \sum_{b=1}^{m} K_{ab} S_{bj} \\
&= \sum_{a=1}^{m} I\{\phi(i) = a\} \sum_{b=1}^{m} K_{ab}\, I\{\phi(j) = b\} \\
&= \sum_{a=1}^{m} I\{\phi(i) = a\} K_{a\phi(j)} \\
&= K_{\phi(i)\phi(j)}.
\end{aligned}$$

The similarity between two vertices is just the similarity between their attributes.

6 Generalization to Arbitrary Number of Attributes

We have thus far assumed that one attribute is observed for each vertex in G, but now we relax that assumption and allow vertices to have an arbitrary number of attributes. Define

$\psi : V \to \mathcal{P}(A)$ to be this mapping from vertices to attribute sets. Then to estimate the sample probabilities of attribute pairs, we may use

$$f_{a,b} = \sum_{\{u,v\} \in E} I\{(a,b) \in \psi(u) \times \psi(v) \text{ or } (a,b) \in \psi(v) \times \psi(u)\},$$

or alternatively,

$$f_{a,b} = \sum_{\{u,v\} \in E} \frac{I\{(a,b) \in \psi(u) \times \psi(v) \text{ or } (a,b) \in \psi(v) \times \psi(u)\}}{|\psi(u)| \cdot |\psi(v)|}.$$

The former method weighs observations of attribute pairs equally, while the latter method weighs edges in the graph equally. Note that either case is problematic when $|\psi(u)| = 0$ for some vertex u: in the first case, edges will simply be ignored whenever one or both of the linked vertices has no attribute, which therefore under-utilizes the information in the graph; in the second case, an even more grievous zero division will occur.

To remedy this, we simply force ψ to map only to non-empty sets of attributes by introducing "dummy attributes" as necessary; that is, we augment the set A with one unique attribute a_u for each node u such that $\psi(u) = \emptyset$, and we let

$$\psi'(u) = \begin{cases} \{a_u\}, & \text{if } \psi(u) = \emptyset \\ \psi(u), & \text{otherwise.} \end{cases}$$

Making this adjustment will allow all vertices and edges to be represented in the pairwise counts. Afterwards we can normalize the $f_{a,b}$ appropriately to turn them into a joint probability distribution, then apply smoothing and/or convert to PMIs as desired.

We can also redefine the collapse operator in the following way. Assuming $\psi : V \to \mathcal{P}(A)$ only maps vertices to non-empty sets, let S be the $m \times n$ matrix where

$$S_{ij} = \frac{I\{i \in \psi(j)\}}{|\psi(j)|}.$$

Then for some similarity kernel K on A, the attribute-induced kernel $K' = S^T K S$ has the property

$$\begin{aligned} K'_{ij} &= \sum_{a=1}^{m} S_{ai} \sum_{b=1}^{m} K_{ab} S_{bj} \\ &= \sum_{a=1}^{m} \frac{I\{a \in \psi(i)\}}{|\psi(i)|} \sum_{b=1}^{m} K_{ab} \frac{I\{b \in \psi(j)\}}{|\psi(j)|} \\ &= \frac{1}{|\psi(i)| \cdot |\psi(j)|} \sum_{a=1}^{m} I\{a \in \psi(i)\} \sum_{b=1}^{m} K_{ab} \, I\{b \in \psi(j)\} \\ &= \frac{1}{|\psi(i)| \cdot |\psi(j)|} \sum_{a \in \psi(i)} \sum_{b \in \psi(j)} K_{ab}. \end{aligned}$$

The similarity between two vertices is the mean of the similarities of all corresponding pairs of attributes.

References

[1] Bergsma, Shane; Dredze, Mark; Van Durme, Benjamin; Wilson, Theresa; and Yarowsky, David. 2013. Broadly Improving User Classification via Communication-Based Name and Location Clustering on Twitter. In *NAACL*.