3.2 Vertex Similarity Algorithm

Let us use the knowledge gained in the previous sections to write a simple vertex similarity algorithm for finding friends' similarity in terms of their common friends. Figure 3.3 shows a simple social graph. Each vertex in the figure represents a person and the edges represent friendship. We want to assign a weight to each edge, representing the similarity between the persons it connects together. The weight of an edge will be higher if the persons it connects, have more common friends. Thus the weight an edge between two persons shows the similarity between the two persons. We use Jaccard Index as the similarity measure which is used for finding similarity between two sets. If A and B are two sets, Jaccard Index can be written as the size of the intersection between A & B divided by the size of the union of A & B:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

For instance, to calculate similarity between Peter and Jack, we can divide the number of common friends between them with the total number of their friends:

$$\begin{aligned} J(Jack, Peter) &= \frac{|\{Peter, Eva\} \cap \{Jack, Tom, Eva\}|}{|\{Peter, Eva\} \cup \{Jack, Tom, Eva\}|} \\ &= \frac{|\{Eva\}|}{|\{Peter, Jack, Tom, Eva\}|} = \frac{1}{4} \end{aligned}$$

Vertex similarity can be calculated in two supersteps:

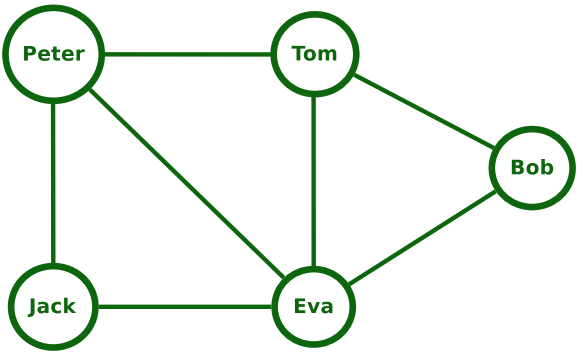

Fig. 3.3 Social graph

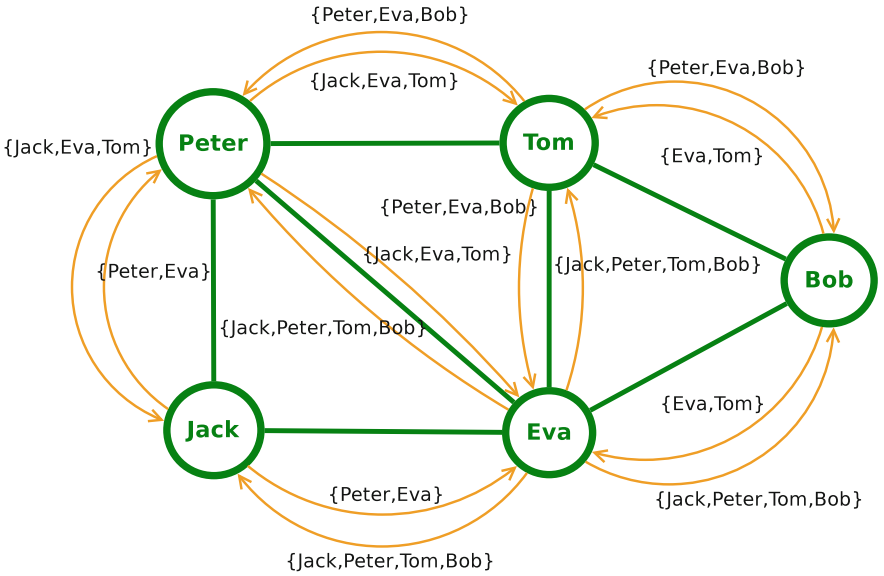

Fig. 3.4 Vertex similarity: superstep 0

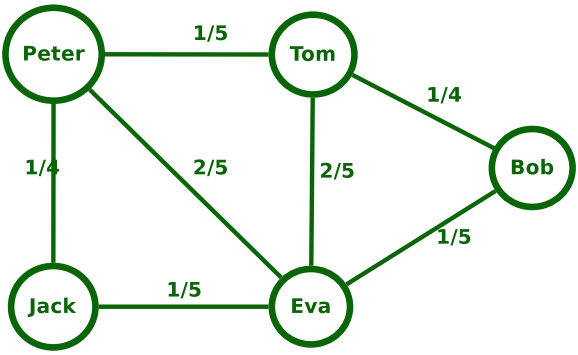

Fig. 3.5 Vertex similarity: superstep 1

1. **Superstep 0**: In the first superstep, each person (vertex) sends the list of his/her friends to each of his/her friends.
2. **Superstep 1**: Each person (vertex) receives the friends's list of his/her friends and calculates its Jaccard similarity with the corresponding friend (Figs. 3.4 and 3.5).

http://www.springer.com/978-3-319-47430-4

Large-Scale Graph Processing Using Apache Giraph

Sakr, S.; Orakzai, F.M.; Abdelaziz, I.; Khayyat, Z.

2016, XXV, 197 p. 102 illus., 87 illus. in color.,
Hardcover
ISBN: 978-3-319-47430-4