14th International Conference on Information Fusion
Chicago, Illinois, USA, July 5-8, 2011

Bow-tie Decomposition in Directed Graphs

Rong Yang
Dept. of Mathematics and Computer Science
Western Kentucky University
Bowling Green, KY 42101, USA
rong.yang@wku.edu

Leyla Zhuhadar and Olfa Nasraoui
Knowledge Discovery and Web Mining Lab
Dept. of Computer Engineering and Computer Science
University of Louisville, KY 40292, USA
leyla.zhuhadar@wku.edu
olfa.nasraoui@louisville.edu

***Abstract*—The bow-tie structure is frequently cited in the literature of the World Wide Web and in many other areas, such as metabolic networks, but it has never been precisely defined, so that to some extent the concept being discussed remains vague. This paper first provides a formal definition of a bow-tie structure relative to a given strongly connected components. That definition details distinctions which are not usually made, such as the difference between intendrils and outtendrils. Theorems and algorithms are then provided to justify and support the definition. Finally a bow-tie decomposition algorithm is developed and illustrated. The algorithms have also been implemented and tested on a university domain.**

Keywords: Graph structure, algorithms, theorems, strongly connected components, bow-tie, social network analysis.

I. Introduction

The web graph is the directed graph which has HTML pages as its nodes and hyperlinks between them as its edges. Understanding the structure of the web graph and its subgraphs is not only of theoretical interest to computer scientists. It can be used to improve vital operations on the web such as browsing, crawling, and community discovery.

Since its introduction by Broder et. al. in [5], the idea of using a bow-tie decomposition as a vehicle for understanding the structure of the world wide web and other directed graphs has been widely used. The original drawing (Figure 1) is a familiar feature in many publications - see [1], [7], [8], [9], [10], [2]. In [7], the bow-tie decomposition, which gives a macroscopic view of the web, served as a jumping-off point for an investigation of finer strucural details within its components. In an investigation of self-similarity in the web, Dill et. al [6] used thematically unified clusters with a bow-tie structure as building blocks to model the web. The evolution of the bow-tie structure over time was studied by Hirate et. al. [9]. Within the application domain, Arasu et. al. [1] carried out computational experiments with the PageRank algorithm and some of its variants using the bow-tie decomposition as the model for the large scale structure of the web. The uneven bow-tie structure of the Java Developer Forum was used in [13] to help test a number of ranking algorithms to identify expertise networks. The bow-tie structure has also shown to be present in core metabolism networks [12]. Thus the bow-tie decomposition has demonstrated its usefulness both in theoretical studies and eminently practical applications.

While the bow-tie structure is frequently cited in the literature, it is generally described in words or by an illustration and has never been precisely defined. It has also never been noted that this structure is really relative to a given strongly connected component. This paper provides the needed formal definition of a bow-tie decomposition relative to a component and an algorithm for computing it.

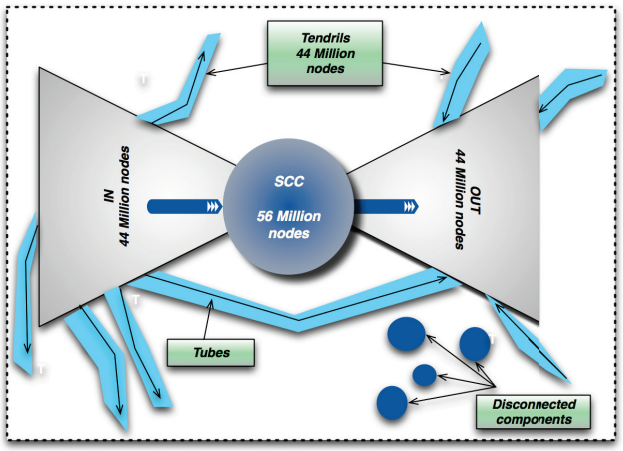

Figure 1. The Original Bowtie Drawing (depicted in [2])

II. Notation and Terminology

Definition 1. A directed graph (or digraph) $G = (V, A)$ consists of a set V of elements called nodes (or vertices), together with a set of ordered pairs of nodes $A \subseteq V \times V$, called arcs. If (v, w) is an arc, v is called the tail of the arc and w is called the head of the arc. The number of nodes, $|V|$, is called the order of the graph and the number of arcs, $|A|$, is called the size of the graph. A graph with order p and size q is often called a (p, q)-graph.

Let $G = (V, A)$ be a digraph and let $v, w \epsilon V$. A path from v to w is a sequence of nodes (distinct except possibly for v and w) $v = v_0, v_1, \ldots, v_n = w$ where for each i = 0, ..., n - 1, (v_i, v_{i+1}), is an arc. This path is of length n (the number of arcs). This is also referred to as a (v, w)-path. If there is such a path, w is said to be reachable from v.

The notion of reachability is extended in the following ways for a node v and subsets of the nodes S and T:

• S is reachable from v if there exists a $w \epsilon S$ such that w is reachable from v.
• v is reachable from S if there exists a $w \epsilon S$ such that v is reachable from w.
• S is reachable from T if there exists a $v \epsilon T$ and a $w \epsilon S$ such that w is reachable from v.

Definition 2. A digraph is said to be strongly connected if every node is reachable from every other node – that is, if any two nodes are mutually reachable.

Given a digraph G, the "mutually reachable" relation is obviously an equivalence relation on the vertices of G. The equivalence classes under this relation are called the strongly connected components of G. They are maximal strongly connected subgraphs of G.

III. The Main Definition

Definition 3. Let $G = (V, A)$ be a digraph and let S be a strongly connected component of G. The bow-tie decomposition of G with respect to S consists of the following sets of nodes:

$$SCC = S$$

$$IN = \{v \epsilon V - S \mid S \, is \, reachable \, from \, v\}$$

$$OUT = \{v \epsilon V - S \mid v \, is \, reachable \, from \, S\}$$

$$TUBES = \{v \epsilon V - S - IN - OUT \mid v \, is \, reachable \, from \, IN \, and \, OUT \, is \, reachable \, from \, v\}$$

$$INTENDRILS = \{v \epsilon V - S \mid v \, is \, reachable \, from \, IN \, and \, OUT \, is \, not \, reachable \, from \, v\}$$

$$OUTTENDRILS = \{v \epsilon V - S \mid v \, is \, not \, reachable \, from \, IN \, and \, OUT \, is \, reachable \, from \, v\}$$

$$OTHERS = V - S - IN - OUT - TUBES - INTENDRILS - OUTTENDRILS$$

This precise definition is in keeping with the original (somewhat informal) definition of the bow-tie structure given in [5]. In particular, there tendrils are described as "containing nodes that are reachable from portions of IN, or that can reach portions of OUT without passage through SCC". No distinction is made in [5] between INTENDRILS and OUTTENDRILS. Also, [5] uses DISCONNECTED where we use OTHERS. DISCONNECTED, as we shall soon see, is not really correct – hence our choice of OTHERS.

The use of the term decomposition in the bow-tie definition is justified by the following:

Theorem 4. *The sets of nodes in the bow-tie decomposition are mutually disjoint and thus form a partition of the nodes.*

Proof: Most of the cases are immediate from the definition. There are only three nontrivial cases. IN and OUT are disjoint, for if they shared a node v in common, then S would be reachable from v by virtue of its being in IN and v would be reachable from S by virtue of its being in OUT. Thus v would have to belong to S, which clearly it does not. IN and INTENDRILS are disjoint because OUT is reachable from any node in IN (via S) while OUT is not reachable from any node in INTENDRILS. OUT and OUTTENDRILS are disjoint because every node in OUT is reachable from any node in IN (via S) while no node in OUTTENDRILS is reachable from IN.

■

A fact which seems to have never appeared in the literature is the following.

Theorem 5. *Each block in the bow-tie decomposition of G with respect to S is the union of strongly connected components of G.*

Proof: Certainly this is true for SCC since by definition it is a strongly connected component of G. IN, OUT, TUBES, INTENDRILS, and OUTTENDRILS are all defined in terms of reachability criteria and any two nodes in the same strongly connected component can reach and can be reached by exactly the same nodes. Thus the assertion holds for these blocks. Finally, OTHERS consists precisely of those nodes which cannot be reached from SCC, IN, or OUT and cannot reach any of those blocks, so the assertion holds for OTHERS as well.

■

The smallest possible digraph in which each of the bow-tie blocks is nonempty has order 7. It is pictured in Figure 2. A slightly larger example (Figure 3) serves to illustrate some of the problems with earlier informal definitions and software. The existing literature does not make the distinction

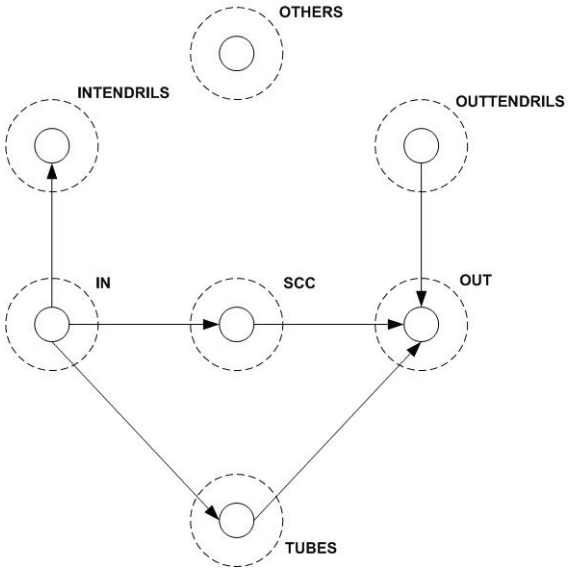

Figure 2. The Smallest Graph With All bow-tie Blocks Nonempty

between intendrils and outtendrils that we are making here. More importantly, when the commonly used social network program Pajek [3], [4], [11] is used to analyze this network, it correctly identifies SCC, IN, OUT, and the TUBES – but it includes nodes 8 and 10 in its collection of tendrils. Only node 9 is placed in the OTHERS category. This is clearly not correct either according to our definition or the original definition in [5].

In the literature about the World Wide Web, reference is typically made to the bow-tie structure of the web or part of the web. What this means is the bow-tie decomposition relative to the strongly connected component of maximum size. Of course, in general there is no such uniquely defined component size. There may be many strongly connected components, all of the same maximum size. Indeed, in the first figure above, each node by itself forms a strongly connected component of maximum size. In practice, there is always a unique maximum size strongly connected component. However, allowing for the bow-tie decomposition to be relative to any chosen strongly connected component does have significant advantages. It may, for example, be used to investigate the portion of the network surrounding any chosen component, say one in IN or OUTTENDRILS, and the resulting finer decomposition can lead to a more detailed understanding of the overall structure. It should be said that our investigations using this new definition shows that there are usually only relatively small strongly connected components within the blocks of the decomposition of web domains (other than SCC itself), so that bow-tie decompositions within the blocks are less likely to be of great interest than might initially be suspected. This is in keeping with the conclusions of [7], where it is suggested that weakly connected components may serve better as tools for analyzing the fine structure of the blocks. Still, it is quite possible that this pattern may not persist over all application areas, so the concept of a bow-tie decomposition relative to a strongly connected component is well worth keeping.

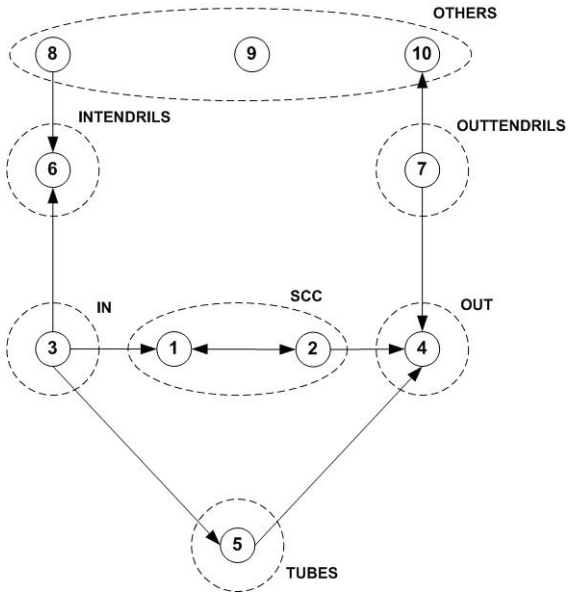

Figure 3. A More Detailed bow-tie

IV. Algorithms

It is easy to find the strongly connected components of a digraph G with the aid of the following definitions.

Definition 6. Let $G = (V, A)$ be a digraph. The transpose of G, G^T, is the digraph formed by using the vertices of G and reversing all of the arcs of G.

If $v \epsilon V$, $DFS_G(v)$ denotes the set of nodes found by a depth-first search in G beginning at v.

It is clear that $DFS_G(v)$ is the set of all nodes that v can reach and that $DFS_{G^T}(v)$ is the set of all nodes that can reach v. The following algorithm is well known and is a standard way of computing strongly connected components. (Despite its slight shortcomings in the bow-tie decomposition, to the best of the authors knowledge, Pajek computes strongly connected components correctly.)

Strongly Connected Component Algorithm: Let $G = (V, A)$ be a digraph and let $v \epsilon V$. Then the strongly connected component containing v is

$$DFS_G(v) \cap DFS_{G^T}(v)$$

The following algorithm for computing the bow-tie decomposition of a digraph with respect to a strongly connected component is a straightforward translation of the

definition.

Bow-tie Decomposition Algorithm: Let $G = (V, A)$ be a digraph and let S be a strongly connected component of G. Then the bow-tie decomposition of G with respect to S may be computed as follows:

1) Set $SCC = S$.
2) Choose $v \epsilon S$. Then $OUT = DFS_G(v) - S$.
3) Choose $v \epsilon S$. Then $IN = DFS_{G^T}(v) - S$.
4) For each $v \epsilon V - S - IN - OUT$, compute the following two Boolean values:

$$IRV = (IN \cap DFS_{G^T}(v) \neq \phi)$$

$$VRO = (OUT \cap DFS_G(v) \neq \phi)$$

Then, since IRV answers the question of whether IN can reach v and VRO answers the question of whether v can reach OUT:

i) IRV and $VRO \Rightarrow v \epsilon TUBES$;

ii) IRV and not $VRO \Rightarrow v \epsilon INTENDRILS$;

iii) not IRV and $VRO \Rightarrow v \epsilon OUTTENDRILS$;

iv) not IRV and not $VRO \Rightarrow v \epsilon OTHER$.

A program based on the above algorithm is relatively efficient, with a runtime of $O\left(|V|^2 + |V||A|\right)$, where $|V|$ is the number of vertices of the graph and $|A|$ is the number of arcs. An unoptimized Java implementation was able to compute the bow-tie structure for the Western Kentucky University domain (26790 nodes, 103131 arcs) in about 40 seconds on a 1.60 GHz machine. The results of that decomposition are given in Table 1.

These numbers are of some interest because of the contrast with the bow-tie structure of the entire Web, where it is estimated that IN, OUT, and TENDRILS are all roughly the same size while SCC is somewhat larger than any of these. They also show that the distinction between INTENDRILS and OUTTENDRILS is a nontrivial one and play significantly different roles within the overall network. Both of these aspects deserve further consideration in future attempts to understand the structure of meaningful portions of the web graph.

Figure 4 shows the actual structure of the WKU domain bow-tie decomposition. The size of the nodes in that graph gives an indication of the size of the respective blocks. Here

Table I
WESTERN KENTUCKY UNIVERSITY BOW-TIE DECOMPOSITION

Region	Size
SCC	*5867*
IN	*104*
OUT	*19250*
TUBES	*2*
INTENDRILS	*446*
OUTTENDRILS	*75*
OTHER	*1046*

it is to be noted that there are links between blocks which are not required by the definition itself - the direct link from IN to OUT, the link from TUBES to INTENDRILS, and so forth. While not required by the definition, neither are they forbidden. Some links are forbidden - there could be no link from OUT to IN, for example.

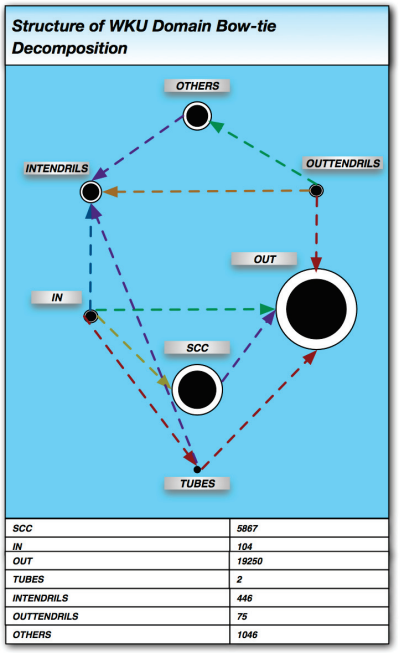

Figure 4. Structure of The WKU Domain bow-tie Decomposition

V. CONCLUSIONS

This paper introduced a precise and unambiguous definition of the bow-tie decomposition for a directed graph which should be useful in standardizing discussions of bowtie structures in the future. It also presented the notion of a bow-tie decomposition relative to a particular strongly connected component which allows for detailed decompositions beyond the gross bow-tie structure. A distinction between INTENDRILS and OUTTENDRILS, previously considered as a single group, was also made, which produces a bowtie decomposition with somewhat finer granularity. Finally, an effective algorithm for producing this decomposition was given along with an example of the type of information it can produce. This work

analyzed the bow-tie structure of the WKU domain based on the proposed new definition and the bow-tie decomposition algorithm. Our future work will consider the comparison study for variety of university domains in Kentucky. In addition, we will conduct a study of relationship between the hubs and authorities within the bow-tie structure.

REFERENCES

[1] A. Arasu, J. Novak, A. Tomkins, and J. Tomlin. PageRank computation and the structure of the web: Experiments and algorithms. In *Proceedings of the Eleventh International World Wide Web Conference, Poster Track*, pages 107–117. Citeseer, 2002.

[2] P. Baldi, P. Frasconi, and P. Smyth. *Modeling the Internet and the Web: Probabilistic methods and algorithms*. John Wiley & Sons Inc, 2003.

[3] V. Batagelj and A. Mrvar. Pajek-program for large network analysis. *Connections*, 21(2):47–57, 1998.

[4] V. Batagelj and A. Mrvar. Pajek analysis and visualization of large networks. In *Graph Drawing*, pages 8–11. Springer, 2002.

[5] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Computer networks*, 33(1-6):309–320, 2000.

[6] S. Dill, R. Kumar, K.S. McCurley, S. Rajagopalan, D. Sivakumar, and A. Tomkins. Self-similarity in the web. *ACM Transactions on Internet Technology (TOIT)*, 2(3):205–223, 2002.

[7] D. Donato, S. Leonardi, S. Millozzi, and P. Tsaparas. Mining the inner structure of the web graph. *Journal of Physics A: Mathematical and Theoretical*, 41:224017, 2008.

[8] D. Easley and J. Kleinberg. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge Univ Pr, 2010.

[9] Y. Hirate, S. Kato, and H. Yamana. Web structure in 2005. *Algorithms and Models for the Web-Graph*, pages 36–46, 2008.

[10] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The web and social networks. *IEEE Computer*, 35(11):32–36, 2002.

[11] W. Nooy, A. Mrvar, and V. Batagelj. Exploratory social network analysis with Pajek. 2005.

[12] R. Tanaka, M. Csete, and J. Doyle. Highly optimised global organisation of metabolic networks. *IEE Proceedings-Systems Biology*, 152(4):179–184, 2005.

[13] J. Zhang, M.S. Ackerman, and L. Adamic. Expertise networks in online communities: structure and algorithms. In *Proceedings of the 16th international conference on World Wide Web*, page 230. ACM, 2007.