

Triangular Line Graphs and Word Sense Disambiguation

Pranav Anand¹, Henry Escudro²,
Raluca Gera³, and Craig Martell⁴

¹Linguistics Department, University of California Santa Cruz,
Santa Cruz, CA 95064; *panand@ucsc.edu*

²Department of Mathematics, Juniata College,
Huntingdon, PA 16652; *escudro@juniata.edu*

³Department of Applied Mathematics, Naval Postgraduate School,
Monterey, CA 93943; *rgera@nps.edu*

⁴Computer Science Department, Naval Postgraduate School,
Monterey, CA 93943; *cmartell@nps.edu*

March 11, 2011

Abstract

Linguists often represent the relationships between words in a collection of text as an undirected graph $G = (V, E)$, where V is the vocabulary and vertices are adjacent in G if and only if the words they represent co-occur in a relevant pattern in the text. Ideally, the words with similar meanings give rise to the vertices of a component of the graph. However, many words have several distinct meanings, preventing components from characterizing distinct semantic fields. This paper examines how the structural properties of triangular line graphs motivate the use of clustering coefficient on the triangular line graph, thereby helping to identify polysemous words. The *triangular line graph* of G , denoted by $T(G)$, is the subgraph of the line graph of G where two vertices are adjacent if the corresponding edges in G belong to a K_3 .

Keywords: H -line Graphs, Triangular Line Graph, Line Graph, Connectivity;

2000 Mathematical subject classification: 05C12, 05C40

1 Introduction and Motivation

One of the chief concerns of linguists is the pervasive ambiguity of natural language. At the lexical (or word) level, this manifests in the existence of the multiplicity of *senses*, or specific meanings, that a word may have. In the hand-compiled Wordnet [5] ontology, for example, 17% of the 114,648 nouns have more than one sense, and the average noun has 1.2 senses. The word ‘lip’, for example, has 5 distinct senses listed in Wordnet, including one from human anatomy (‘either of two fleshy folds of tissue that surround the mouth and play a role in speaking’) and one describing an object’s structure (‘the top edge of a vessel or other container’).

The multiplicity of word senses considerably complicates computational tasks over language – including automatic translation [14], information retrieval [12], and speech synthesis [13] – and has led to an interest in automatic *word sense disambiguation* (WSD), which determines the word sense intended in a particular context (see [6] and [9] for a broader history and overview of WSD). In [10] and [11], the authors independently observe that certain linguistic patterns, such as coordination patterns ‘ w_1 and w_2 ’, are highly predictive if w_1 and w_2 have senses in the same semantic field. Thus, observing ‘lip and mouth’ and ‘lip and nose’ would allow a system to discover that there are senses of ‘lip’, ‘mouth’, and ‘nose’ which are semantically related. A natural representation for these word relationships is an undirected graph $G = (V, E)$, where V is the vocabulary (the set of distinct words in the text) and vertices are adjacent in G if and only if the words they represent co-occur in a relevant pattern in the text. Ideally, the words in the same semantic field will give rise to the vertices of a component of the graph. However, given that ‘lip’ has multiple senses, such a system also runs the risk of placing ‘handle’ and ‘teeth’ in the same component. What is needed is a method for spotting these spurious links.

In [3], Dorow *et al.* proposed an algorithm based on triangles in a graph. They argue that while a word w may co-occur with word w_1 under sense S_1 and with w_2 under sense S_2 (for example, *lip* with *handle* and with *nose*), it is unlikely that all three words co-occur with each other; that is, w, w_1, w_2 form a triangle in G . Thus, each component should be disconnected at any vertex v with neighbors v', v'' in separate components of $\langle N(v) \rangle$. To visualize these vertices, Dorow *et al.* introduced the *link graph*, which is equivalent to the anti-Gallai graph of G (see Le [8]) or triangular line graph of G (see Dorrough [4] and Jarrett [7]), denoted by $T(G)$. In this paper we use the terminology of triangular line graphs. $T(G)$ is the subgraph of the line graph of G where two adjacent vertices in $T(G)$ correspond to two edges that belong to a K_3 in G . $T(G)$ is itself an instance of the H -line graph introduced by Chartrand, Gavlas and Schultz in [1] for $H \cong K_3$.

We evaluated the triangular line graph operator on a graph constructed from the English Gigaword corpus, which consists of 1 billion words of English text.¹ This procedure produced 460 components with half of the words in one component. Thus, the triangular line graph procedure must be supplemented when used in a large corpus. This paper explores the structural properties of the triangular line graph, with the aim of understanding how best to use such graphs for word sense disambiguation. We will demonstrate that the properties of the triangular line graph, particularly those of K_n , allow us to effectively bound the clustering coefficient metric, thereby assisting in disambiguating ambiguous words.

2 Definitions and Observations

In this paper, all graphs $G = (V(G), E(G))$ are simple graphs (no multiple edges), with vertex set $V(G)$ and edge set $E(G)$. The order of G is $|V(G)|$ while the size of G is $|E(G)|$. For graph theory terminology and notation used in this paper, we refer the reader to [2]. We first recall the definition of triangular line graph as defined by Jarrett in [7].

¹In this paper, we use the model of [3], where each noun is represented by a vertex, and an edge is present between two vertices if the corresponding nouns are separated by either “and”, “or”, or a comma.

Definition 2.1. The triangular line graph $T(G)$ of a graph G is the graph with vertex set $E(G)$, where two distinct vertices e and f are adjacent in $T(G)$ if and only if there exists a subgraph $H \cong K_3$ of G with $e, f \in E(H)$. Any such subgraph is called a triangle of G .

For $\ell \geq 1$, we use the notation of [4] where $\Gamma_\ell(G)$ denotes the number of subgraphs of G isomorphic to K_ℓ . The following two observations appeared in [4] and [7] and are useful for this paper.

Observation 2.2.

- (a) If H is a subgraph of G , then $T(H)$ is a subgraph of $T(G)$.
- (b) Let G be a graph. If e is an edge of G that does not belong to a copy of K_3 in G , then e is an isolated vertex in $T(G)$.
- (c) If G is a graph, then every vertex in $T(G)$ is either an isolated vertex or belongs to a copy of K_3 in $T(G)$.

Observation 2.3. If G is a graph, then

- (a) the order of $T(G)$ is equal to the size of G ; that is, $\Gamma_1(T(G)) = \Gamma_2(G)$, and
- (b) the size of $T(G)$ is triple the number of triangles in G ; that is, $\Gamma_2(T(G)) = 3\Gamma_3(G)$.

We now present general results about the triangular line graph of a given graph.

Observation 2.4. For any graph G , every vertex in $T(G)$ has even degree.

Corollary 2.5. For a graph G , the triangular line graph $T(G)$ is Eulerian if and only if $T(G)$ is connected.

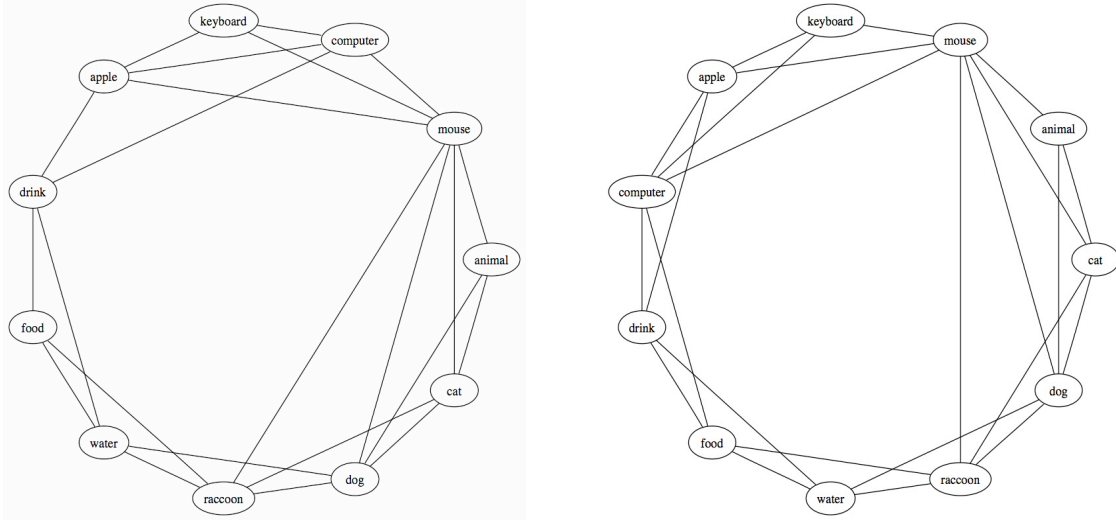
This brings up the question of which triangular line graphs are connected. Note that the triangular line graph of a connected graph does not have to be connected. For example, the triangular line graph of the bow tie graph $2K_2 + K_1$ is $2K_3$ which is disconnected.

For a positive integer n , we define a *triangle trail* to be a graph that consist of n copies of K_3 , say $\Delta_1, \Delta_2, \dots, \Delta_n$, where Δ_i and Δ_{i+1} share a common edge, say $e_{i,i+1}$. If no Δ_i ($1 \leq i \leq n$) is repeated, the triangle trail is said to be a *triangle path*.

To illustrate what a triangle trail is, consider the two graphs of Figure 1. The first graph does not have a triangle trail joining the triangle $\langle \text{drink}, \text{apple}, \text{computer} \rangle$ to the triangle $\langle \text{drink}, \text{food}, \text{water} \rangle$ while the second one has a triangle trail between any two triangles.

We say that a connected graph G belongs to *class C* if (1) every edge of G belongs to a K_3 , and (2) for every two copies of K_3 in G , there is a triangle trail that connects them.

Theorem 2.6. For any graph G , the triangular line graph $T(G)$ is connected if and only if G belongs to class C .



(a) A graph which does not have a triangle trail going from the triangle $\langle \text{drink}, \text{apple}, \text{computer} \rangle$ to the triangle $\langle \text{drink}, \text{food}, \text{water} \rangle$.

(b) A graph in which each triangle is connected to any other triangle via a triangle path.

Figure 1: Two possible collocation graphs for the same nouns

Proof. Suppose first that $T(G)$ is connected. If G has an edge that does not belong to a K_3 , then by Observation 2.2, the corresponding vertex in $T(G)$ will be an isolated vertex which contradicts the fact that $T(G)$ is connected. Hence, every edge in G belongs to a K_3 . Now suppose that Δ and Δ' are two copies of K_3 in G , ($n > 1$). Let $e \in E(\Delta)$ and $e' \in E(\Delta')$ such that the distance between the vertices of e and e' is a minimum. Since $T(G)$ is connected, the corresponding vertices v_e and $v_{e'}$ are connected by a path in $T(G)$, say $v_e = v_{e_1}, v_{e_2}, \dots, v_{e_n} = v_{e'}$. Since, by Observation 2.2, every edge of $T(G)$ belongs to a triangle, it follows that for each edge $v_{e_i}v_{e_{i+1}}$ there is a triangle in $T(G)$ that contains both vertices. It follows that the edges e_i and e_{i+1} belong to a common triangle in G , say $\Delta_{i,i+1}$. Since $\Delta_{i-1,i}$ and $\Delta_{i,i+1}$ share the edge e_i , we have a triangle path that connects Δ and Δ' in G .

For the converse, we assume that G is connected. Let v_e and $v_{e'}$ be two vertices in $T(G)$ corresponding to the two edges e and e' in G . Since every edge in $T(G)$ belongs to a K_3 , it follows that $e \in \Delta$ and $e' \in \Delta'$ where Δ and Δ' are triangles in G . Since G is in class \mathcal{C} , there is a triangle trail that connects Δ and Δ' , say $\Delta = \Delta_1, \Delta_2, \dots, \Delta_n = \Delta'$. This implies that there is an edge trail $e - e_{1,2} - e_{2,3} - \dots - e_{n-1,n} - e'$ that connects e and e' , where $e_{i,i+1}$ is an edge common to Δ_i and Δ_{i+1} . Thus, in $T(G)$ we have the corresponding vertex trail $v_e, v_{e_{1,2}}, v_{e_{2,3}}, \dots, v_{e_{n-1,n}}, v_{e'}$ connecting v_{e_i} and v_{e_j} . Thus $T(G)$ is connected. ■

One standard question is to determine which graphs are isomorphic to their triangular line graphs; that is, to find the fixed points of the operator that sends a graph to its triangular line graph. It is easily proved that in the case of line graphs, this only happens for cycles. In [4], it is demonstrated that the fixed point of the triangular line graph operator for an arbitrary graph G is isomorphic to the disjoint union of $r \geq 0$ triangles. From this, it follows that a graph is isomorphic to itself if and only if it is a disjoint union of triangles.

3 Recognition of Triangular Line Graphs

In this section, we will consider the following question: Can a given graph H be the triangular line graph of some graph? For example, the path P_n is not the triangular line graph of any graph since a triangular line graph is either empty or has girth 3. Also, every edge of a triangular line graph must belong to a K_3 . Thus, if a graph H has an edge that does not belong to a K_3 , then H is not a triangular line graph.

In answering this question, we may restrict our attention to connected graphs, as any component of a triangular line graph is a triangular line graph as well.

Proposition 3.1. *If G is a graph, then $T(G)$ contains a copy of $K_4 - e$ if and only if G contains a copy of K_4 .*

Proof. Let G be a graph. Suppose first that $T(G)$ has copy of $K_4 - e$ whose vertices are u, v, w and x , where u and x are not necessarily adjacent. Let e_u, e_v, e_w and e_x be the edges in G that correspond to u, v, w and x , respectively. It follows that each of the following pairs of edges in G lies in a copy of K_3 in G : $\{e_u, e_v\}, \{e_u, e_w\}, \{e_x, e_v\}, \{e_x, e_w\}$ and $\{e_v, e_w\}$ (and, $\{e_u, e_x\}$ if $K_4 - e$ is not induced). Observe that e_v and e_w must be incident to a common vertex in G while their non-common vertices must be adjacent. Since each of e_u and e_x appears with each of e_v and e_w in the pairs of edges listed above, at least one of e_u and e_x (say e_u) must be incident to the vertex that is common to e_v and e_w . Moreover, the non-common vertices of e_u and e_v are adjacent in G as well as the non-common vertices of e_u and e_w . This implies that the vertices of e_u, e_v and e_w induce a copy of K_4 in G .

Suppose now that G contains a copy of K_4 . Since $T(K_4)$ contains a copy of $K_4 - e$, it follows that $T(G)$ contains a copy of $K_4 - e$. ■

Corollary 3.2. *If H is a graph that contains a copy of $K_4 - e$ and $H = T(G)$ for some graph G , then H contains a copy of the octahedral graph.*

Proof. By Proposition 3.1, it follows that K_4 is a subgraph of G . This implies that $T(K_4)$ (which is the octahedral graph) is a subgraph of $T(G) = H$ by Observation 2.2. ■

Let $W_n = K_1 + C_n$ (also know as an n -wheel) where $n \geq 3$. The vertex that dominates the vertices on the n -cycle of W_n is the *center* of W_n and the edges incident to the center are *spokes*. For $n \geq 4$, the triangular line graph of W_n is called an n -cog-wheel and is denoted by CW_n . The n -cog-wheel CW_n is the graph obtained from an n -cycle $C : v_0, v_1, \dots, v_{n-1}, v_0$ by introducing n new vertices u_0, u_1, \dots, u_{n-1} , and then joining u_i to v_i and v_{i+1} for each i where addition is done modulo n . Two cog-wheels CW_n and CW_m are *non-overlapping* if $V(CW_n) \cap V(CW_m) = \emptyset$.

Note that two triangles in a graph may either be edge-adjacent (like $K_4 - e$), vertex-adjacent (like $2K_2 + K_1$), or disjoint. Let class \mathcal{D} be the set of all graphs G such that: (a) every edge of G belongs to a copy of K_3 , (b) G does not contain $(K_4 - e)$ as a subgraph and (c) all the cog-wheels in G are non-overlapping.

Proposition 3.3. *Every graph in \mathcal{D} is a triangular line graph.*

Proof. Let $H \in \mathcal{D}$ where $V(H) = \{v_1, v_2, \dots, v_n\}$. Since every edge in H belongs a triangle, it follows that every vertex in H belongs to a triangle also. We construct a preimage G as follows. For each triangle $\langle v_x, v_y, v_z \rangle$ in H , construct a triangle T_{xyz} and label its edges x, y and z . Let $\mathcal{T} = \{\Delta_1, \Delta_2, \dots, \Delta_{\Gamma_3(H)}\}$ be the set of all such triangles T_{xyz} and let $\mathcal{C} = \{CW_{m_1}, \dots, CW_{m_w}\}$ be the set of all cog-wheels in H . Note that if there exist two triangles in H that have a common vertex v_x , then there are two distinct triangles $\Delta_i, \Delta_j \in \mathcal{T}$ such that each of Δ_i and Δ_j has an edge labeled x and all the other edges of Δ_i and Δ_j have different labels. For each CW_n in \mathcal{C} , construct a copy of W_n by identifying the edges having the same labels in the corresponding triangles in the appropriate way – the edges that correspond to the vertices on the n -cycle of CW_n will be the spokes of W_n while the rest of the edges will lie on the n -cycle of W_n . Denote the set of such wheels by \mathcal{W} , and \mathcal{T}' the set of all triangles that belong to a wheel in \mathcal{W} . From the set $\mathcal{W} \cup (\mathcal{T} - \mathcal{T}')$, construct the graph G by identifying all edges having the same labels.

We now show that $T(G) \cong H$. Suppose $\Delta_i, \Delta_j, \Delta_k \in \mathcal{T}$ form a K_4 such that $\Delta_i = T_{abc}, \Delta_j = T_{ade}$ and $\Delta_k = T_{bdf}$. Then $\langle v_a, v_b, v_c \rangle, \langle v_a, v_d, v_e \rangle$ and $\langle v_a, v_d, v_f \rangle$ are triangles in H . However, this implies that $K_4 - e$ is a subgraph of H which is a contradiction. Hence, if Δ_i, Δ_j and Δ_k are distinct triangles in \mathcal{T} , then at least one of the pairs $\{\Delta_i, \Delta_j\}, \{\Delta_i, \Delta_k\}$ and $\{\Delta_j, \Delta_k\}$ have edges all of which have different labels. From this and by the way G was constructed, G does not have K_4 as a subgraph. Hence, $K_4 - e$ is not a subgraph of $T(G)$. Observe that the only way a triangle can be formed in G that does not correspond to a triangle in H is to identify the edges labeled x of two triangles $\langle e_1, e_2, x \rangle$ and $\langle f_1, f_2, x \rangle$ for which the vertex incident to both e_1 and e_2 (call this vertex v) is adjacent to the vertex incident to both f_1 and f_2 (call this vertex w). Now the edge vw must then belong to a triangle Δ , with Δ sharing only the vertex v with triangle $\langle e_1, e_2, x \rangle$, which is contrary to our construction of G . Thus, every triangle in G correspond to a triangle in H . Since no two cog-wheels overlap in H , we also know that when two triangles in \mathcal{T} are joined by identifying edges with the same label in the construction of G , no edge label is lost. Thus, $E(G) = \{1, 2, \dots, n\}$ and so $|V(H)| = |E(G)| = |V(T(G))|$. Let $V(T(G)) = \{u_1, u_2, \dots, u_n\}$ and consider the mapping $\phi : V(T(G)) \rightarrow V(H)$ given by $\phi(u_i) = v_i$ for $i = 1, 2, \dots, n$. Since ϕ is bijective, it suffices to show that ϕ preserves adjacency. Suppose $u_x u_y \in E(T(G))$. It follows that the edges x and y belong to a common triangle Δ in G . Let z be the other edge of this triangle. This means that there is a triangle $T_{xyz} \in \mathcal{T}$ that was obtained from some triangle $\langle v_x, v_y, v_z \rangle$ in H . Hence, $v_x v_y \in E(H)$. Using a similar argument, we can show that if $u_x u_y \notin E(T(G))$, then $v_x v_y \notin E(H)$. Thus, $T(G) \cong H$. ■

Proposition 3.4. *The complete graph K_n ($n \geq 4$) is not the triangular line graph of any graph.*

Proof. Assume, to the contrary, that there is a graph G such that $T(G) = K_n$ for some $n \geq 4$. It follows that G has at least four edges say e_1, e_2, e_3 and e_4 . Moreover, each pair of edges in G occurs in a triangle. Without loss of generality, we assume that $\langle e_1, e_2, e_3 \rangle$ is a triangle. Observe that e_4 cannot share endpoints with all of e_1, e_2 and e_3 . Hence, e_4 cannot occur in a triangle with each of e_1, e_2 and e_3 and we get a contradiction. ■

Note that K_2 is not the triangular line graph of any graph either. Thus, the only complete graphs that are triangular graphs of some graph are K_1 and K_3 .

In summary, if H is the triangular line graph of some graph, then the following must be true: a) Every vertex in H has even degree; b) The size of H is a multiple of 3; c) Every edge in H belongs to a triangle; d) If $K_4 - e$ is a subgraph of H , then the octahedral graph is a subgraph of H . The converse of this statement is not true. For example, K_7 satisfies properties (a) through (d) but is not a triangular line graph by Proposition 3.4.

4 The Clustering Coefficient and the Triangular Line Graph

In this section, we use the structure of the triangular line graph of the complete graph K_n to decide how to best use the clustering coefficient to identify polysemous words. Since $T(K_2) = K_1$ and $T(K_3) = K_3$, we focus our attention to the case when $n \geq 4$. If $V(K_n) = \{1, 2, \dots, n\}$ is the vertex set of the complete graph K_n , then the vertex set of the triangular line graph of K_n is given by $V(T(K_n)) = \{v_{ij} : 1 \leq i \neq j \leq n\}$.

Now let i be a vertex in K_n . Since every pair of edges in K_n belongs to a copy of K_3 in K_n , it follows that the set of edges $\{ij : 1 \leq j \leq n, j \neq i\}$ gives rise to a clique of order $n - 1$ in $T(K_n)$. As in [7], we denote each of the above mentioned cliques by G_i for $i = 1, 2, \dots, n$. Note that if $1 \leq j, k, \ell \leq n$ and j, k and ℓ are distinct, then the triangle in $T(K_n)$ induced by the vertices $jk, j\ell$ and $k\ell$ does not belong to any of the cliques G_i mentioned above. We call these triangles *clique-linking triangles* and denote them by $T_{j,k,\ell}$. Since each clique-linking triangle $T_{j,k,\ell}$ is determined by the vertices j, k and ℓ in K_n , it follows that $T(K_n)$ has $\binom{n}{3}$ clique-linking triangles. Also, from [4] we have that $T(K_n)$ has a total of $\binom{n}{3} + 4\binom{n}{4}$ triangles. We incorporate this formally in the following result which describes the other structural characteristics of the triangular line graph of the complete graph K_n , ($n \geq 4$), based on the results in [4] and [7].

Theorem 4.1. *Let $n \geq 4$ be an integer.*

1. $T(K_n)$ is connected, $2(n - 2)$ -regular of order $\binom{n}{2}$ and size $\binom{n}{2}(n - 2)$.
2. $T(K_n)$ has exactly n distinct copies of K_{n-1} .
3. $T(K_n)$ has $\binom{n}{3} + 4\binom{n}{4}$ triangles, $\binom{n}{3}$ of which are clique-linking triangles.
4. For every $i \neq j$ with $1 \leq i, j \leq n$, there is a unique vertex common to G_i and G_j , namely ij .
5. The subgraph induced by the set of vertices $V(G_i) \cup V(G_j)$, where $i \neq j$ and $1 \leq i, j \leq n$, contains $n - 2$ clique-linking triangles, namely those of the type $T_{i,j,k}$ where $k \notin \{i, j\}$, $1 \leq k \leq n$.
6. For every distinct triple of integers i, j, k with $1 \leq i, j, k \leq n$, there is a unique clique-linking triangle that joins all of G_i, G_j and G_k together.

We now return to the problem of decomposing the triangular line graph mentioned at the outset of the paper. In [3], the authors considered applying the clustering coefficient (or curvature) of a vertex to identify polysemous words. The curvature of a vertex w , $curv(w)$ was defined in [3] as

$$curv(w) = \frac{\text{number of triangles } w \text{ participates in}}{\text{number of triangles } w \text{ could participate in}}.$$

and is thus a measure of the completeness of a neighborhood. Note that the number of triangles that a vertex w can participate in is given by $|N(w)||N(w) - 1|/2$. We propose that curvature, known in graph theory as *clustering coefficient*, be used on the triangular line graph instead of the graph itself. In particular, we identify vertices with low clustering coefficient in $T(G)$ whose neighborhood contains vertices with high clustering coefficient. Vertices with high clustering coefficient in $T(G)$ correspond to word pairs whose neighbors are highly interconnected, and thus are part of a semantically homogenous group of senses. Vertices with low clustering coefficient in $T(G)$ correspond to word pairs whose neighbors are loosely connected, and hence appear infrequently. We are interested in removing erroneous links – those links which we observe in a corpus but which are not semantically meaningful. Given what was said above, these correspond to vertices with low clustering coefficient in the neighborhood of vertices with high clustering coefficient – cases where a polysemous word’s two senses are connected because of a small, semantically meaningless cases of co-occurrence in the corpus.

To see this, first note that the clustering coefficient of a vertex in $T(G)$ is 1 if and only if the vertex has degree 2 in $T(G)$. The edges represented by these vertices in G can be safely removed in the attempt of identifying the ambiguous words in the text. Figure 2 shows the triangular line graph of the graph in Figure 1(b). Notice the vertices with low clustering coefficient adjacent to vertices of degree 2.

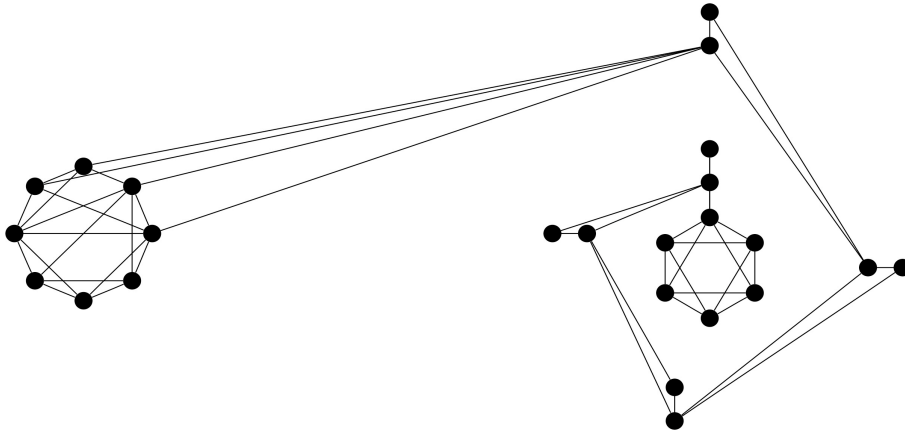
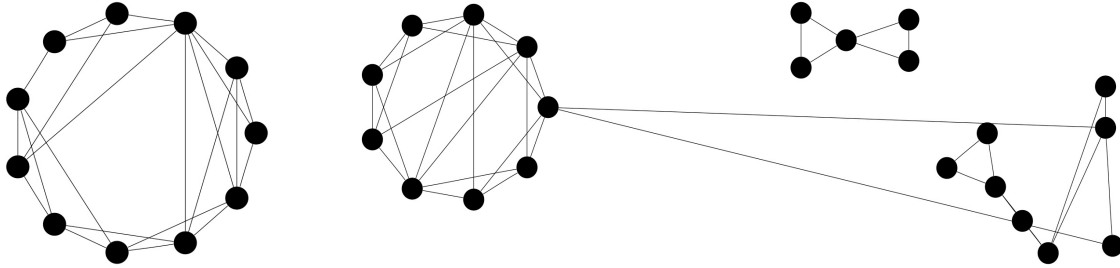


Figure 2: The connected triangular line graph of the graph in 1(b).

If the edges corresponding to vertices with low clustering coefficient in $T(G)$ (like *drink-apple* in our case) were removed from G to create G^* , then $T(G^*)$ would separate the semantic meanings of the words (like the meanings of the word *apple*, see Figure 3).

Thus, we suggest that vertices with low clustering coefficient are likely indications of polysemy when adjacent to vertices with high clustering coefficient. From Theorem 4.1, we have the following intrinsic lower bound on the clustering coefficient of a vertex of $T(K_n)$, where $n \geq 3$.



(a) The graph in Figure 1(b) with the edge drink-apple removed.

(b) The triangular line graph of the graph on the left

Figure 3: Each component presents a different semantic space once the edge with the lowest clustering coefficient was removed from the graph in 1(b)

Corollary 4.2. *If $n \geq 3$ is an integer, then the clustering coefficient of v is greater than $1/2$ for all $v \in T(K_n)$.*

This suggests that we consider the clustering coefficient of a vertex v in $T(G)$ to be high if $curv(v)$ is greater than $1/2$. Since some sets of words that belong to the same semantic field do not give rise to cliques, it is reasonable to consider the clustering coefficient of vertices v in $T(G)$ to be high if $curv(v)$ is slightly less than $1/2$. Our evaluation of this method with high clustering coefficient defined as above 0.3 and low clustering coefficient defined as below 0.05 resulted in the removal of 15% of the edges in the original graph from the Gigaword corpus. Application of the triangular line graph transformation to this new trimmed graph resulted in 900 components, with the largest containing only 5% of the words in the data set, a major division of what occurred without this procedure. Recall that the triangular line graph of the Gigaword corpus graph had one component with over 50% of the words, and thus was not effective in distinguishing the meanings of the words.

5 Conclusion and Remarks

This paper investigated the basic graph-theoretic properties of triangular line graphs, and their application to automate the discovery of ambiguous words. The more general concept of H -line graph was introduced by Chartrand *et al* [1] while triangular line graphs were studied by Jarrett [7] and Dorrough [4] who studied the convergence of sequences of iterated triangular graphs for the complete graph and for a general graph, respectively. In addition to discussing connectedness and vertex degrees in triangular line graphs, we also identified a large family of graphs whose members are triangular line graphs, and presented characteristics for potential triangular line graphs.

The description of the triangular line graph of the complete graph K_n in [7], allows one to compute an upper bound on a threshold for an encouraging procedure to split triangular line graphs using clustering coefficients. The procedure builds on a re-estimation component common in the machine-learning literature on hill-climbing, suggesting the fruitfulness of further study on iterative applications of the triangular line graph transformation and how such transformations help in word sense disambiguation.

Acknowledgement: The authors would like to thank the referees, as well as Derrick Stolee and Stephen Hartke, for their comments and suggestions regarding this publication. Also P. Anand, R. Gera and C. Martell thank the Department of the Navy PEO Integrated Warfare Systems for funding that partially supported this work.

References

- [1] G. Chartrand, H.J. Gavlas and M. Schultz, *Convergence of sequences of iterated H-line graphs*, Discrete Mathematics 147 (1995) 73–86.
- [2] G. Chartrand and P. Zhang, *Introduction to Graph Theory*, McGraw-Hill, Kalamazoo, MI (2004).
- [3] B. Dorow, D. Widdows, K Ling, J.-P. Eckmann, D. Sergi and E. Moses, *Using Curvature and Markov Clustering in Graphs for Lexical Acquisition and Word Sense Discrimination*, CSLI Publications, Stanford, CA (2004).
- [4] D. Dorrough, *Convergence of sequences of iterated triangular line graphs*, Discrete Mathematics 161 (1996) 79–86.
- [5] C. Fellbaum, *Wordnet: An Electronic Lexical Database*, MIT Press, Cambridge, MA (1998).
- [6] N. Ide, Nancy and J. Véronis, *Word Sense Disambiguation: The State of the Art*, Computational Linguistics 24(1) (1998) 1-41.
- [7] E.B. Jarrett, *On iterated triangular line graphs*, Proceedings : 7th International Kalamazoo Conference in Graph Theory, Combinatorics, Algorithms, and Applications, Wiley, NY (1995) 589–599.
- [8] Van Bang Le, *Gallai graphs and anti-Gallai graphs*, Discrete Mathematics 159, (1996) 179–189.
- [9] R. Mihalcea and T. Pedersen, *Advances in Word Sense Disambiguation*, AAAI Tutorial, July 9, 2005., <http://www.d.umn.edu/~tpederse/Tutorials/ADVANCES-IN-WSD-AAAI-2005.ppt> (2005).
- [10] E. Riloff and J. Shepherd, *A corpus based approach for building semantic lexicons*, Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing, EMNLP-97, (1997) 117-124.
- [11] B. Roark and E. Charniak, *Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction*, Proceedings of the 17th International Conference on Computational Linguistics/38th Annual Meeting for the Computational Linguistics, COLING-ACL-98, (1998) 1110-1116.
- [12] G. Salton, A. Wong, C.S. and Yang,(1975), *A vector space for information retrieval*, Communications of the ACM 18(11) (1975) 613-620.
- [13] R. Sproat, J. Hirschberg and D. Yarowsky, *A corpus-based synthesizer*, Proceedings of the International Conference on Spoken Language Processing, Banff, Alberta, Canada, (1992).

- [14] W. Weaver (1949), Translation. Mimeographed, 12 pp., July 15, 1949. Reprinted in W. N. Locke and D. A. Booth, (1955) (Eds.), *Machine translation of languages*. John Wiley & Sons, New York, 15-23. New Mexico State University, Las Cruces, New Mexico.