Structure based Graph Distance Measures of High Degree of Precision

Yanghua Xiao[1,∗], Hua Dong[2], Wentao Wu[1], Momiao Xiong[2,3], Wei Wang[1], Baile Shi[1]

[1]*Department of Computing and Information Technology, Fudan University, Shanghai, China*

[2] *Theoretical Systems Biology Lab , School of Life Science, Fudan University, Shanghai, China*

[3]*Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX 77225, USA*

Abstract

In recent years, evaluating graph distance has become more and more important in a variety of real applications and many graph distance measures have been proposed. Among all of those measures, structure-based graph distance has become the research focus due to its independence of the definition of cost function. However, the existing structure-based graph distance measures have low degree of precision because only node and edge information of graphs are employed in these graphs metrics. To improve the precision of graph distance measure, we define a substructure abundance vector (SAV) to capture more substructure information of a graph. Furthermore, based on the SAV, we propose unified graph distance measures which are generalization of the existing structurebased graph distance measures. In general, the unified graph distance measures can evaluate graph distance in much finer grain. We also show that unified graph distance measures based on occurrence mapping and some of their variants are metrics. Finally, we apply the unified graph distance metric and its variants to the population evolution analysis and construct distance graphs of marker networks in three populations, which reflect the single nucleotide polymorphism (SNP) linkage disequilibrium (LD) differences among these populations.

Keywords: graph distance; distance metric; structure-based graph distance; SNP linkage disequilibrium

1 Introduction

As a data structure, graph has been widely used to represent un-structured data, model complex interaction relations among objects and define concepts. Compared to other data structures such as sequence, tree, graph is more sophisticated and more general, and consequently studies on graph have been a research hotspot.

Many real applications usually need to measure the similarity or distance between objects represented by graph. For example, in computer ?visualization and pattern recognition [26], similarity between unknown graph pattern and model graph pattern must be measured in the well known graph matching process. In chemical study, similarity searching based on 2D representation of molecular structure is one of the most common approaches to virtual screening [6, 12], where in some cases, appropriate measure of inter-molecular structural similarity is the key of the searching task.

Therefore, it is of great interest to measure the graph distance or similarity in various applications [23, 25, 28]. Great efforts have been devoted to studying graph distance measures in different application fields over the past decades. As a result, various graph distance measures have been proposed in the literatures [1, 5, 6, 19, 24, 26, 27]. These graph distance measures can be classified into three classes: *cost-based distance measures*, *structure-based distance measures* and *feature-based distance measures*. In [13], cost-based distance and structure-based distance are considered as one class, because it has been proved in [2] that given certain cost functions, the structure-based

[∗] Corresponding author. Tel +86 21 55075013, P.O.Box:200433

email: shawyanghua@gmail.com (Yanghua Xiao), hdong0425@gmail.com (Hua Dong)

graph distance measures, such as graph distance measures based upon Maximal Common Subgraph (MCS) [1] [*], are equivalent to corresponding edit distance measures with certain cost functions.

In pattern recognition, considering error tolerance or error correcting, many cost-based graph distances [21, 22] have been proposed, which are measured by the minimum edit cost to transform one graph into another. However, due to the complexity in selection of cost functions, many recently published works [5, 6, 13, 23] adopted the structure-based distance measures, which do not rely on the cost functions. In structure-based distance measures, the common substructure or superstructure has been considered as the measure of the degree of the similarity between graph patterns. Besides these two kinds of distance measures, feature-based measures have also been widely studied in chemical-informatics and bio-informatics. In feature-based measures, distance or similarity has been measured according to the feature vectors derived from the chemical or biological structures. Hence, the effects of the feature-based measures heavily rely on the definition of the characteristic structures.

Due to the presence of some effective algorithms [6] and the independence of cost functions or characteristic structures, structure-based measures, especially those measures based on maximal common subgraph have become the most popular graph distance measures in recent years.

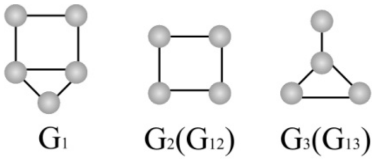


Figure 1: Three graphs G_1, G_2, G_3 and two maximal common subgraph G_{12}, G_{13}

Although various structure-based graph distance or similarity measures have been available, many graph pairs in some application domains can not be correctly measured using these measures. For example, as shown in Figure 1, given three graphs G_1, G_2 and G_3, we need to evaluate the similarity or distance among these graphs. If MCES-based distance metric, a widely used graph distance metric, is used, the maximal common subgraph G_{12} (between G_1 and G_2) and the maximum common subgraph G_{13} (between G_1 and G_3) will have the same number of nodes and edges. Consequently, we can reach the conclusion that G_2 is similar to G_1 to the same extent as G_3 similar to G_1.

However, in the following sections, we will show that G_{13} contains much richer substructure information than G_{12}. As shown in Figure 2, G_{13} contains some unique substructures, such as *triangle* and *star*, which do not appear in G_{12}. Hence, from such *substructure abundance* perspective, G_{13} is intuitively of more significance than G_{12}; and consequently, G_3 should be evaluated to be more similar to G_1 than G_2 to G_1. Therefore, the *richness of the unique substructures* occurring in a graph can contribute to the evaluation of graph distance, which is the basic principle underlying the measures we proposed in this paper.

Since nodes and edges are elementary constituents of a graph, size about nodes or edges in maximal common subgraphs will be a significant indication of the similarity between graphs, which is the fundamental idea of existing structure-based graph distance measures. For example, two representatives of them, MCIS-based graph distance[1] and MCES-based graph distance[12] use the number of *nodes* of MCIS, and the number of the *edges* of MCES, *respectively,* to evaluate the similarity between two graphs. However, in our studies, besides node or edge information in maximal common subgraph, information about more complex and larger substructures that will occur in maximal common subgraph will be utilized to evaluate distance between a graph pair.

[*] The term ‘Maximal Common Subgraph (MCS)’ has been widely used, but it also has brought much confusion to the existing literatures. Strictly speaking, the graph distance metric proposed in [1] is based on *maximal common vertex induced subgraph*, abbreviated as MCIS , and some following graph distance metrics are based on *maximum common edge induced graph*, abbreviated as MCES. In this paper, to distinguish these two concepts, we will explicitly use MCIS or MCES, instead of MCS.

In the following parts , we will show that structural differences between graphs can be amplified when considering information of larger substructures. Thus, if we evaluate graph distance in terms of certain larger or more complex substructure instead of some trivial substructures, such as nodes or edges, we can evaluate graph distance with higher degree of precision or in much finer grain than graph distance measures based on MCIS or MCES.

Evaluating graph distance according to *richness of the unique substructures* is also practically meaningful in many real applications. For example, in the analysis protein-protein interaction , protein-DNA and gene-gene interaction networks, it has been widely believed that substructures of these networks represent certain *functional modules* of cells or organisms. Thus, in Figure 1, if *triangle* and *star* appearing in G_{13} are considered as functional modules of biological networks, then G_{13} will contain more abundant functional modules than G_{12}. Consequently, we can naturally come to the conclusion that G_3 is more similar to G_1 than G_2 to G_1. Hence, comparing protein networks in terms of substructure information is biologically meaningful.

To accurately quantify graph distance is in great demand for many applications, especially for researches on evolution of biology networks. For example, we could use Bayesian Networks [10] to study SNPs [8] LD structure and their evolutions among different populations [18]. In such studies, how to measure similarity or distance among the constructed networks is an interesting but challenging problem. One of the great challenges is that traditional MCS-based graph distance metric can only evaluate the graph distance in much coarser grain, which can not satisfy the requirement of identifying the minute difference between different population structures. Hence, it’s of great need to devise new graph distance measures that can evaluate graph distances precisely.

2 Preliminaries

We begin this section with some basic notations. Let $G= (V,E,L,l)$ be a *labeled graph*, where V is the set of vertices, E is the set of edges and $E \subseteq V \times V$, L is the set of labels, and $l: V \cup E \rightarrow L$ is a labeling function that assigns a label to an edge or a vertex. Note that graph labeling is one of key issues in problems related to graph isomorphism. However, in some contexts, where graph isomorphism is not significant, *G also can be denoted as a 2-tuple (V, E)*.

The vertex set of G is referred to as $V(G)$, and its edge set as $E(G)$. A *path P* in a graph is a sequence of vertices $v_1, v_2, \ldots, v_k$, where $v_i \in V$ and $v_i v_{i+1} \in E$. The vertices v_1 and v_k are linked by P and are called the *ends* of path P. The number of edges of a path is its *length*, and the path of length k is denoted as P^k. A path is *simple* if its vertices are all distinct. A graph G is called *connected* if for any vertices $u, v \in V(G)$, there exist a path with ends u, v. A graph $G=(V,E)$ is called *subgraph* of $G'=(V', E')$, denoted as $G \subseteq G'$, if and only if $E \subseteq E'$ and $V \subseteq V'$. If graph $G=(V,E)$ is a subgraph of $G'=(V', E')$ such that $E=V \times V \cap (E')$， then G is a *vertex induced subgraph* of G', in the contexts without confusions, it is often called as *induced subgraph.* If graph $G(V,E)$ is a subgraph of G' such that $V=V(E)$，then G is an *edge induced subgraph* of G' . Obviously, as an edge induced subgraph, it will contain many isolated nodes, which are often considered as trivial in many real applications.

Definition 2.1 (Graph isomorphism). Given graphs $G = (V, E, L, l)$ and $G' = (V', E', L', l')$. A *bijective* function f : $V \rightarrow V'$ is called a *graph isomorphism* from G to G' if (1) for any vertex $u \in V$, $l(u) = l'(f(u))$; (2) for any edge $(u, v) \in E$, we have $(f(u), f(v)) \in E'$ and $l(u, v) = l'(f(u), f(v))$; for any edge $(u', v') \in E'$, $(f^{-1}(u'), f^{-1}(v')) \in E$ and $l'(u', v') = l(f^{-1}(u'), f^{-1}(v'))$.

Definition 2.2 (Subgraph isomorphism). An *injective* function f : $V \rightarrow V'$ is a *subgraph isomorphism* from $G = (V, E, L, l)$ to $G' = (V', E', L', l')$, if there exists a subgraph $S \subseteq G'$ such that f is a graph isomorphism from G to S.

If there exists a graph isomorphism between G and G',we call G is *isomorphic* to G', and denoted as $G \cong G'$. If there exists a subgraph isomorphism from G to G', we call G *subgraph isomorphic* to G', and denoted as $G \leqq G'$.

Graph isomorphism and subgraph isomorphism are two essential concepts to describe relations between graphs, which underlie the study of the whole graph space. Hence, we first need to gain deeper insight into the properties of these two graph relations, which are described by the following two propositions that are immediate consequences of the definitions.

Proposition 2.1: *Graph isomorphism between graphs is an equivalence relation***.**

Proposition 2.2: *Subgraph isomorphism relation between graphs is transitive.*

Given a class of graphs, we can define measures on graphs, such as the number of nodes of a graph, the diameter of a graph and so on. In real applications, we expect that the two isomorphic graphs have the same values under certain measure on graphs. Graph measures satisfying such desired properties are referred to as *vertex invariants*, which are formally defined as follows:

Definition 2.3 (Graph Invariant). Let $\boldsymbol{G}$ be the set of graphs，f: $\boldsymbol{G} \rightarrow \boldsymbol{R}^{\rho}$ is called a *(ρ -dimensional) graph invariant* if $G \cong G' \Rightarrow f(G)=f(G')$. If $f(G)=f(G') \Rightarrow G \cong G'$ is also true, then f is called a *complete graph invariant*.

A graph G_{12} is a *common edge induced subgraph* of G_1 and G_2, if G_{12} is isomorphic to edge induced subgraphs of G_1 and G_2, respectively. A *maximum*[*] *common edge subgraph* (MCES) is a common edge induced subgraph of G_1 and G_2 with the largest number of edges. Without explicit statements, in the following discussions, MCS always indicates MCES.

In many real applications, it is desired that the graph distance measures possess certain properties. For example, one may wish that the distance from graph G_1 to G_2 is the same as that from G_2 to G_1. Generally speaking, it is often desired that a distance measure fulfill the properties of a *metric*, which is defined in Definition 2.4. But in some cases, the properties listed in Definition 2.4 are too restrictive, or incompatible with the problem domain under consideration.

Definition 2.4(Graph Distance Metric) Let $\boldsymbol{G}$ be the set of graphs , the mapping d: $\boldsymbol{G} \times \boldsymbol{G} \rightarrow \boldsymbol{R}$ is called a graph distance metric，if $\forall G_1, G_2, G_3 \in \boldsymbol{G}$，the following properties hold true:

(1) $d(G_1,G_2) \geq 0$ (non-negativity)
(2) $d(G_1,G_2)=0 \Leftrightarrow G_1 \cong G_2$ (uniqueness)
(3) $d(G_1,G_2)=d(G_2,G_1)$ (symmetry)
(4) $d(G_1,G_2)+ d(G_2,G_3) \geq d(G_1,G_3)$ (triangle inequality)

And the ordered pair ($\boldsymbol{G}$, d) is a *metric space*.

In some specifications, uniqueness is equivalent to other two properties: *positiveness* and *reflexivity*. $d(G_1,G_2)=0 \Rightarrow G_1 \cong G_2$ is called as *positiveness* , because it is equivalent to that $\forall G_1, G_2 \in \boldsymbol{G}$, if G_1 is not isomorphic to G_2 , $d(G_1,G_2)>0$. $G_1 \cong G_2 \Rightarrow d(G_1,G_2)=0$ is referred to as *reflexivity*. If *positiveness* does not hold for d, then d is a *pseudo-metric* and ($\boldsymbol{G}$, d) is a *pseudo-metric space*. Obviously, *pseudo-metric* space is a generalization of a metric space in which we allow the possibility that $d(G_1,G_2)=0$ for non-isomorphic graphs G_1 and G_2.

Strictly speaking, the uniqueness of a graph distance measure only holds , when *isomorphic* graphs can be considered as *equal*. But this assumption is certainly justified in most applications [1]. Another issue that needs to be addressed is that *positiveness* is usually too restrictive in real applications. As a result, many graph distance

[*] Generally speaking, given a class of common graphs of G_1 and G_2, denoted as $\boldsymbol{G}=\{g_1,g_2,...,g_n\}$, 'maximum' corresponds to a linear order defined on $\boldsymbol{G}$ according to the size of each common graph, 'maximal' corresponds to a partial order defined on $\boldsymbol{G}$ according to '$\subseteq$' or '≦'relation between graphs.

measures in real applications are only *pseudo-metrics*.

3 Structure Abundance Vector

Despite the importance of substructure information of graphs, no existing mathematic concepts can be utilized to describe them appropriately. In this section, we propose a new concept: Structure Abundance Vector, to capture the substructure information of a graph.

Given a labeled graph $G=(V,E,L,l)$, let $S(G)=\{g|g\leqq G\}$ be the set consisting of graphs that are subgraph isomorphic to G. Since graph isomorphic relation is an equivalent relation on graphs, we can obtain a *quotient set* of $S(G)$ *w.r.t* graph isomorphism relation ($\cong$). Such quotient set can be denoted as $S(G)/\cong=\{[g_1],\ldots,[g_n]\}$ with $[g_i]=\{g|g\in S(G),g\cong g_i\}$ for each $1\leq i\leq n$, where $[g_i]$ represents an equivalent class *w.r.t* graph isomorphic relation and g_i is the representative of the equivalence class. We call $[g_i]$ a *pattern in G*, and each graph belonging to $[g_i]$ is called a *pattern graph*. Among these pattern graphs, those occurring in *G, i.e.* those subgraphs of *G*, are called *occurrences in G of pattern* $[g_i]$.

Generally speaking, in many application domains, not all pattern graphs of $[g_i]$ but those *occurrences in G of pattern* $[g_i]$ are of interests. Hence, without loss of generality, we can select one of occurrences in *G* of pattern $[g_i]$ to represent the pattern. In such way, we obtain a set $\Gamma(G)=\{g_1, \ldots, g_n\}$ *s.t.* $\forall g_i, g_j\subseteq G\ (i\neq j)$, g_i is not *isomorphic* to g_j. In other words, $\Gamma(G)$ consists of all subgraphs (subpatterns) of *G* that are *non-isomorphic* to each other.

However, in some cases, different occurrences of the same pattern do make sense, we have to make an alternative choice. In these cases, we may define $\Gamma(G)$ to be the set consisting of all the *G*'s subgraphs(subpatterns) that are not *equal* to each other, *i.e.*, $\Gamma(G)=\{g_1, \ldots, g_m\}$ *s.t.* for any two subgraphs $g_i,g_j\subseteq G(i\neq j)$, $g_i\neq g_i$.

Furthermore, $\Gamma(G)$ can be partitioned according to the size of the subgraphs, here we use the number of edges to quantify the size of the graph. Thus, $\Gamma(G)$ can be partitioned into $\{\Gamma(G)_1, \Gamma(G)_2\ldots\Gamma(G)_m\}$ ($m=|E(G)|$) with $\Gamma(G)_i$ representing the subset of $\Gamma(G)$ in which each graph has i edges. Naturally, $\Gamma(G)$ and $\Gamma(G)_i$ can be associated with corresponding mappings, in the context without confusions denoted as Γ and Γ_i, which map each graph to its subgraphs or subpatterns (with size i). Thus, we get $\Gamma(G)=\Gamma_1\ (G)\cup\ldots\cup\Gamma_m(G)$, and we refer to each Γ_i as a *substructure mapping* of a graph. Since $\Gamma(G)$ can be defined as the pattern set or occurrence set, we need to further subdivide *substructure mappings* into two elementary classes, one is *pattern mapping* corresponding to the non-isomorphic patterns, the other is *occurrence mapping* corresponding to the non-equal occurrence. The formal definition is given as follows.

Definition 3.1(Pattern Mapping): A *pattern mapping* Γ_i is a substructure mapping such that for every graph *G*, $\Gamma_i(G)(\ 0\leq i\leq|E(G)|)$ is the set of *G*'s edge-induced subgraphs with *i* edges and any two graphs in $\Gamma_i(G)$ are *non-isomorphic* to each other.

Definition 3.2(Occurrence Mapping): An *occurrence mapping* Γ_i is a substructure mapping such that for every graph *G*, $\Gamma_i(G)(\ 0\leq i\leq|E(G)|)$ is the set of *G*'s edge-induced subgraphs with *i* edges and any two graphs in $\Gamma_i(G)$ are *non-equal* to each other.

Pleasenote that in the above definitions, *i* may equal to 0. In this case, edge-induced subgraphs with 0 edges indicate to nodes in a graph; and consequently $\Gamma_0(G)$ represent the node set of the graph. In the following discussion, without explicit statements, $\Gamma_0(G)$ always represents the node set of graph *G*.

Let $\Gamma=\{\Gamma_i|0\leq i\leq|E(G)|\ \}$ be the set of all pattern mappings or occurrence mappings for graph *G*, then we can define a measure on graph *G* to summarize the information of substructures in *G* according to the substructure-mapping set

Γ. Such measure can be easily defined as a vector: $\vec{V}=(|\Gamma_0(G)|,\ldots,|\Gamma_m(G)|)$, where denotes the number of elements in the set Γ, $m=|E(G)|$. Obviously, the vector expresses the abundance of the substructures of a graph *G* in terms of the size of the substructure, so we call this vector a *structure abundance vector* of graph G.

Definition 3.3 (SAV: *Structure Abundance Vector*). A structure abundance vector of a graph *G* is an $(|E(G)|+1)$-dimensional vector, whose *i*-th ($0\le i\le |E(G)|$) dimension is the number of the *G*'s edge-induced subgraphs with *i* edges such that these graphs not isomorphic/equal to each other.

Theorem 3.1 .*Structure Abundance Vector is a graph invariant.*

It's easy to prove that if $G\cong G'$, we have $\vec{V}(G)=\vec{V}(G')$. Hence, $\vec{V}$ is a graph invariant.

Example 3.1: As shown in Figure 2, G_2 and G_3 have the same number of vertices and edges, while G_3 has richer non-isomorphic substructures, especially, in columnΓ_3. The structure abundance can be evaluated by $\vec{V}$ in terms of pattern mapping. Thus we have $\vec{V}(G_2)$=(1,1,2,1,1), $\vec{V}(G_3)$=(1,1,2,3,1). Note that if the focus of the problem domain is not non-isomorphic patterns but non-equal occurrences of different patterns. We have $\vec{V}(G_2)$=(4,4,6,4,1), $\vec{V}(G_3)$=(4,4,6,4,1).

Note that if structure abundance vector is defined in terms of *occurrence mappings,* the vector can be computed directly by $\vec{V}(G)=(n, C_m^1, C_m^2, \ldots, C_m^m)$, where $n=|V(G)|$ and $m=|E(G)|$.

In some real applications, disconnected substructures are often treated as trivial substructure or as noisy data. Therefore, in these applications, it is necessary to take into account the *connectivity constraint* of substructures to exclude those disconnected substructures. Thus, in Example 3.1, if the substructure mapping Γ_i is restricted to obtain only those connected substructures, then the disconnected substructures that are marked with dotted line in Figure 2 will be discarded. Thus, when Γ_i is pattern mapping, we have $\vec{V}(G_2)$=(1,1,1,1,1) and $\vec{V}(G_3)$=(1,1,1,3,1); when Γ_i is occurrence mapping, we have $\vec{V}(G_2)$=(4,4,4,4,1), $\vec{V}(G_3)$=(4,4,5,4,1).

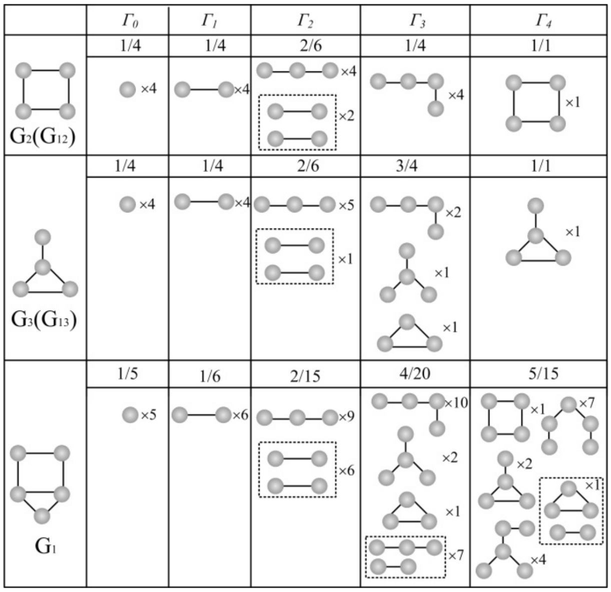


Figure 2: Substructures in G_1, G_2 (G_{12}), and G_3 (G_{13})

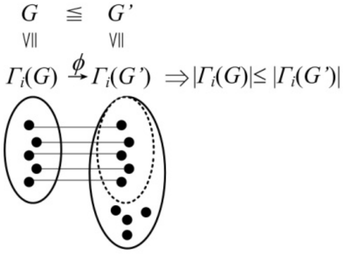

G ≦ G'
VII
VII
$\Gamma_i(G) \xrightarrow{\phi} \Gamma_i(G') \Rightarrow |\Gamma_i(G)| \le |\Gamma_i(G')|$

Figure 3: Relation among *G, G', $\Gamma_i(G)$* and *$\Gamma_i(G')$*

4 Graph Distance Measures based on SAV

In this section we will first discuss graph relationship under substructure mapping, which is essential for study of the distance measures based on SAV. Before the detailed discussion, we first give some basic notations. Let $\boldsymbol{G}$ be the set of all distinct labeled graphs. Given two labeled graph G_1 (V_1,E_1,L_1,l_1) and G_2 (V_2,E_2,L_2,l_2) belonging to $\boldsymbol{G}$, let $G_{12}=mces(G_1,G_2)$, and $\Gamma=\{\Gamma_i|0\leqslant i\leqslant|E(G_{12})|\}$.

4.1 Relations between Graphs under Substructure Mapping

In the following discussion, it's necessary to extend '≦' from relation between graphs to relation between graph sets. For this purpose, we first define a property of any given graph with respect to '≦' relation between graph sets.

Property 4.1: Let $\boldsymbol{H}$ be a labeled graphs set, a mapping $p_{\boldsymbol{H}}$: $\boldsymbol{G}\rightarrow\{0,1\}$ is a property of graphs, which is defined in the way that $p_{\boldsymbol{H}}$ $(G\in\boldsymbol{G})=1$ *if* $\forall g\in\boldsymbol{H}$, $g\leqq G$; otherwise, $p_{\boldsymbol{H}}$ $(G\in\boldsymbol{G})=0$.

Lemma 4.1: Given a set of labeled graphs $\boldsymbol{H}$, if $p_{\boldsymbol{H}}$ $(G)=1$, then for any $G\leqq G'$, we have $p_{\boldsymbol{H}}$ $(G')=1$.

Proof: From $p_{\boldsymbol{H}}(G)=1$, we have $\forall g\in\boldsymbol{H}$, $g\leqq G$. Since '≦' relation between graphs is transitive (Proposition 2.2), it follows naturally that $\forall g\in\boldsymbol{H}$, $g\leqq G'$. Thus, we have $p_{\boldsymbol{H}}$ $(G')=1$. □

We can denote the statement that $\forall g\in\boldsymbol{H}$, $g\leqq G$ by $\boldsymbol{H}\leqq G$. Similarly, the statement that $\forall g\in\boldsymbol{H}$, $\forall g'\in\boldsymbol{G}$, $g\leqq g'$ also can be denoted by $\boldsymbol{H}\leqq\boldsymbol{G}$. Obviously, transitive property of relation '≦' also holds for graph sets. Based on extended graph relation '≦', we can further study the relation between graphs under substructure mapping, which is stated in Theorem 4.1.

Theorem 4.1: Given a pattern mapping Γ_i, for any two graphs $G\leqq G'$, the following statements hold:

(1) There is an *injective mapping* ϕ: $\Gamma_i(G)\rightarrow\Gamma_i(G')$. for each $g\in\Gamma_i(G)$, there is only one unique $\phi(g)\in\Gamma_i(G')$ *s.t.* $g\cong\phi(g)$.

(2) $|\Gamma_i(G)|\leq|\Gamma_i(G')|$

(3) $|\Gamma_i(G)|=|\Gamma_i(G')|$ if $G\cong G'$.

Proof: Since $\forall g\in\Gamma_i(G)$, $g\leqq G$, we have $\forall g\in\Gamma_i(G)$, $g\leqq G\leqq G'$. Thus, for each $g\in\Gamma_i(G)$, there exists a unique $g'\in\Gamma_i(G')$ *s.t.* $g'\cong g$. Furthermore $\forall g_1,g_2\in\Gamma_i(G)$, if $g_1\neq g_2$, we have $g_1'\neq g_2'$, where $g_1', g_2'\in\Gamma_i(G')$ and $g_1'\cong g_1$, $g_2'\cong g_2$ (Note that since Γ_i is a pattern mapping, then for $\forall g_1,g_2\in\Gamma_i(G_1)$, $g_1\neq g_2$ also implies that g_1 and g_2 are non-isomorphic to each other). Hence, we can construct an injective mapping ϕ from $\Gamma_i(G)$ to $\Gamma_i(G')$, as described in statement (1).

It follows directly from statement (1) that $|\Gamma_i(G)|\leq|\Gamma_i(G')|$. When $G\cong G'$, $|\Gamma_i(G)|=|\Gamma_i(G')|$ and the mapping ϕ: $\Gamma_i(G)\rightarrow\Gamma_i(G')$ is *surjective, i.e.* for each $g'\in\Gamma_i(G')$ there is some $g\in\Gamma_i(G)$ *s.t.* $\phi(g)=g'$. Hence ϕ: $\Gamma_i(G)\rightarrow\Gamma_i(G')$ is *bijective* or *one-to-one correspondence,* when $G\cong G'$. The relation among G, G', $\Gamma_i(G)$ and $\Gamma_i(G')$ that described in Theorem 4.1 is shown in Figure 3.

Note that in Theorem 4.1, if pattern mapping Γ_i is replaced by an occurrence mapping, all the statements still hold. Furthermore, statement (3) can be replaced with a stronger assertion, which is described in Theorem 4.2. Hence, to prove Theorem 4.2, we only need to show $|\Gamma_i(G)|=|\Gamma_i(G')|\Rightarrow G\cong G'$. $|\Gamma_i(G)|=|\Gamma_i(G')|$ implies that $C_m^i=C_{m'}^i$ ($m=|E(G)|$ and $m'=|E(G')|$), so $m=m'$. Since $G\leqq G'$, we have $G\cong G'$.

Theorem 4.2: Given an occurrence mapping Γ_i, then for any two graphs $G\leqq G'$, the following statements hold:

(1) There exists an *injective mapping* ϕ: $\Gamma_i(G)\rightarrow\Gamma_i(G')$ such that for each $g\in\Gamma_i(G)$, there is only one unique $\phi(g)\in\Gamma_i(G')$ *s.t.* $g\cong\phi(g)$.

(2) $|\Gamma_i(G)|\leq|\Gamma_i(G')|$

(3) $|\Gamma_i(G)| = |\Gamma_i(G')|$ if and only if $G ≌ G'$.

Corollary 4.1: Given a substructure mapping (occurrence mapping or pattern mapping) Γ_i, for any three graphs G_1, G_2 and G , if $G_1 ≦ G$, $G_2 ≦ G$ and $\forall g_1 \in \Gamma_i(G_1)$, $\forall g_2 \in \Gamma_i(G_2)$, g_1 is not isomorphic to g_2, then the following statements hold:

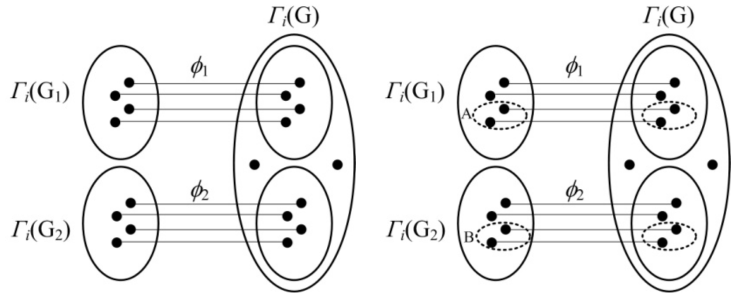
(1)There exist *injective mappings* ϕ_1: $\Gamma_i(G_1) \rightarrow \Gamma_i(G)$ and ϕ_2: $\Gamma_i(G_2) \rightarrow \Gamma_i(G)$ such that $\phi_1(\Gamma_i(G_1)) \cap \phi_2(\Gamma_i(G_2)) = \varnothing$ and $\phi_1(\Gamma_i(G_1)) \cup \phi_2(\Gamma_i(G_2)) \subseteq \Gamma_i(G)$.

(2) $|\Gamma_i(G_1)| + |\Gamma_i(G_2)| \leq |\Gamma_i(G)|$

Proof: Since $G_1 ≦ G$, according to Theorem 4.1and 4.2, there must exist an injective mapping ϕ_1: $\Gamma_i(G_1) \rightarrow \Gamma_i(G)$ *s.t.* $\forall g \in \Gamma_i(G_1)$, $\phi_1(g) \in \Gamma_i(G)$ and $g ≌ \phi_1(g)$.Similarly, there must exist an injective mapping ϕ_2: $\Gamma_i(G_2) \rightarrow \Gamma_i(G)$ *s.t.* $\forall g \in \Gamma_i(G_2)$, $\phi_2(g) \in \Gamma_i(G)$ and $g ≌ \phi_2(g)$. Obviously, $\phi_1(\Gamma_i(G_1)) \subseteq \Gamma_i(G)$, $\phi_2(\Gamma_i(G_2)) \subseteq \Gamma_i(G)$, so $\phi_1(\Gamma_i(G_1)) \cup \phi_2(\Gamma_i(G_2)) \subseteq \Gamma_i(G)$ Hence, to prove the statement (1) hold true, we only need to show that $\phi_1(\Gamma_i(G_1)) \cap \phi_2(\Gamma_i(G_2)) = \varnothing$.

Assume $\phi_1(\Gamma_i(G_1)) \cap \phi_2(\Gamma_i(G_2)) \neq \varnothing$, there must exist $g \in \Gamma_i(G)$ such that $\phi_1^{-1}(g) \in \Gamma_i(G_1)$,$\phi_2^{-1}(g) \in \Gamma_i(G_2)$ and $g ≌ \phi_1^{-1}(g) ≌ \phi_2^{-1}(g)$, which contradict to the known condition that $\forall g_1 \in \Gamma_i(G_1)$, $\forall g_2 \in \Gamma_i(G_2)$, g_1 is not isomorphic to g_2.

Statement (2) can be directly inferred from Statement (1). The mapping relations of $\Gamma_i(G_1)$, $\Gamma_i(G_2)$ and $\Gamma_i(G)$ are shown in Figure 4(a). □



(a) (b)

Figure 4 Illustration of proof procedure of Corollary 1 and 2

An immediate consequence of Corollary 4.1 is the following Corollary 4.2. The detailed proof of Corollary 4.2 is similar to that of Corollary 1 and is omitted in this paper. The illustration of the proof procedure is shown in Figure 4(b).

Corollary 4.2: Given an substructure mapping (occurrence mapping or pattern mapping) Γ_i, for any three graphs G_1, G_2 and G , $G_1 ≦ G$, $G_2 ≦ G$, let $A \subseteq \Gamma_i(G_1)$ and $B \subseteq \Gamma_i(G_2)$, if $\forall g_1 \in A$, $\forall g_2 \in B$, g_1 is not isomorphic to g_2, then $|A| + |B| \leq |\Gamma_i(G)|$.

4.2 Unified Graph Distance Measures based on SAV

All the existing structure-based graph distance measures can be expressed in the common form: $d(G_1,G_2) = 1 - m(G_{12})/M(G_1,G_2)$, with $m(G_1,G_2)$ representing the similarity of graphs and $M(G_1,G_2)$ representing the size of the problem. Generally, $M(G_1,G_2)$ can be defined in the following three cases:

Case 1: $max(|\Gamma_i(G_1)|, |\Gamma_i(G_2)|)$;

Case 2: $min(|\Gamma_i(G_1)|, |\Gamma_i(G_2)|)$;

Case 3: $|\Gamma_i(G_1)| + |\Gamma_i(G_2)| - |\Gamma_i(G_{12})|$;

Following this common form, we can give two elementary graph distance measures that are based on substructure abundance of graphs.

Definition 4.1. The distance of two non-empty graphs G_1 and G_2 is defined as $d_i(G_1,G_2) = 1 - |\Gamma_i(G_{12})|/M(|\Gamma_i(G_1)|,$

$|\mathit{\Gamma}_i(G_2)|)$, where $\mathit{\Gamma}_i$ is a *pattern mapping* with $i \le |E(G_{12})|$[*] and $M(|\mathit{\Gamma}_i(G_1)|,|\mathit{\Gamma}_i(G_2)|)$ is defined as one of Case 1,2,3.

Definition 4.2. The distance of two non-empty graphs G_1 and G_2 is defined as $d_i(G_1,G_2)=1-|\mathit{\Gamma}_i(G_{12})|/M(|\mathit{\Gamma}_i(G_1)|, |\mathit{\Gamma}_i(G_2)|)$, where $\mathit{\Gamma}_i$ is an *occurrence mapping* with $i \le |E(G_{12})|$ and $M(|\mathit{\Gamma}_i(G_1)|,|\mathit{\Gamma}_i(G_2)|)$ is defined as one of Case 1,2,3.

Example 4.1: We shall continue Example 3.1. Let $\mathit{\Gamma}_3$ be a pattern mapping and $M(|\mathit{\Gamma}_3(G_1)|,|\mathit{\Gamma}_3(G_2)|)= max(|\mathit{\Gamma}_3(G_1)|,|\mathit{\Gamma}_3(G_2)|)$, then $d_3(G_1,G_2)=1-\mathit{\Gamma}_3(G_{12})/max(|\mathit{\Gamma}_3(G_1)|,|\mathit{\Gamma}_3(G_2)|)=1-1/max(4,1)=3/4$. Similarly, we have $d_3(G_1,G_3)=1-\mathit{\Gamma}_3(G_{13})/ \quad max(|\mathit{\Gamma}_3(G_1)|,|\mathit{\Gamma}_3(G_3)|)=1-3/max(4,3)=1/4$; $d_3(G_2,G_3)=1-\mathit{\Gamma}_3(G_{23})/max(|\mathit{\Gamma}_3(G_2)|,|\mathit{\Gamma}_3(G_3)|)=1-1/max(1,3)=2/3$.

Let $\mathit{\Gamma}_3$ be an occurrence mapping , then $d_3(G_1,G_2)= d_3(G_1,G_3)=1- C_4^3/max(C_6^3, C_4^3)=1-4/max(20,4)=4/5$. Similarly, we have $d_3(G_2,G_3)=1- C_3^3/ max(C_4^3, C_4^3)=1-1/max(4,4)=3/4$.

Theorem 4.3. For any graphs G_1,G_2 and G_3, the following properties hold true for graph distance measure defined in Definition 4.1, (1)Non-negativity, (2) Reflexivity,(3)Symmetry,(4)Triangle Inequality

Proof: We only give the proof for graph distance measure that is defined in Case 1. The proofs in Case 2 and 3 are similar to the proof in Case 1. In the remaining part of the paper, without explicit statements, all the proof is given for graph distance measure defined in Case 1.

1. Non-negativity.

From Theorem 4.1, it follows that $|\mathit{\Gamma}_i(G_{12})| \le |\mathit{\Gamma}_i(G_1)|$ and $|\mathit{\Gamma}_i(G_{12})| \le |\mathit{\Gamma}_i(G_2)|$, which implies that $|\mathit{\Gamma}_i(G_{12})| \le max (|\mathit{\Gamma}_i(G_1)|, |\mathit{\Gamma}_i(G_2)|)$.

2. Reflexivity.

Recall that structure abundance vector is a graph invariant, which is shown in Theorem 3.1. Thus for any two isomorphic graphs $G_1 \cong G_2$, $\vec{V}(G_1)=\vec{V}(G_2)$ and $G_{12} \cong G_1 \cong G_2$. Consequently, the i-th dimensions of the vector of G_{12}, G_1, G_2 are equal, *i.e.* $|\mathit{\Gamma}_i(G_1)| =|\mathit{\Gamma}_i(G_2)|= |\mathit{\Gamma}_i(G_{12})|$. Hence $G_1 \cong G_2 \Rightarrow d(G_1,G_2)=0$.

3. Symmetry

It follows directly from the definition of the graph distance measure.

4. Triangle Inequality

The detailed proof of triangle inequality is shown in Appendix A. □

Theorem 4.4. For any graphs ., the following properties hold true for graph distance measure defined in Definition 4.2, (1)Non-negativity, (2)Uniqueness,(3)Symmetry,(4)Triangle Inequality

Proof: We only need to show that $d(G_1,G_2)=0 \Rightarrow G_1 \cong G_2$. The proof of other properties is the same as the proof of corresponding properties in Theorem 4.3.

$d(G_1,G_2)=0 \Rightarrow |\mathit{\Gamma}_i(G_{12})| =max(|\mathit{\Gamma}_i(G_1)|, |\mathit{\Gamma}_i(G_2)|)$. Assume $|\mathit{\Gamma}_i(G_1)| \ge |\mathit{\Gamma}_i(G_2)|$, then $|\mathit{\Gamma}_i(G_1)|= |\mathit{\Gamma}_i(G_{12})|$. Since $G_{12} \leqq G_1$, then from statement 2 of Theorem 4.2, we have $G_{12} \cong G_1$. Similarly, by the assumption $|\mathit{\Gamma}_i(G_1)| \ge |\mathit{\Gamma}_i(G_2)|$, we have $|\mathit{\Gamma}_i(G_{12})| =|\mathit{\Gamma}_i(G_1)| \ge |\mathit{\Gamma}_i(G_2)|$. On the other hand, from $G_{12} \leqq G_2$, we have $|\mathit{\Gamma}_i(G_{12})| \le |\mathit{\Gamma}_i(G_2)|$. Hence, we get $|\mathit{\Gamma}_i(G_{12})| =|\mathit{\Gamma}_i(G_2)|$ and $G_{12} \cong G_2$. Thus it follows that $G_{12} \cong G_1 \cong G_2$. So $d(G_1,G_2)=0 \Rightarrow G_1 \cong G_2$.

Through Theorem 4.3, we show that *graph distance measure defined in terms of pattern mapping is a pseudo-metric*. As a pseudo-metric, graph distance measure based on pattern mapping possesses most properties of a metric except for uniqueness, which implies that we can not determine whether two graphs are isomorphic solely given the information that the distance between them is zero. Through Theorem 4.4, we show that *graph distance*

[*] In general, when considering problems of evaluating distance among a class of graphs, for example $\mathbf{G}=\{G_1,G_2 ,\dots,G_n\}$ $(n \ge 2)$, the following condition supposed to be satisfied $i \le min(|E(G_1|,\dots,|E(G_n)|)$, otherwise $M(|\mathit{\Gamma}_i(G_1)|, |\mathit{\Gamma}_i(G_2)|)$ may be zero. We also can let $i \le min(|E(G_{ij})|)_{1 \le i,j \le n}$, which is a stronger condition and make $|\mathit{\Gamma}_i(G_{ij})|$ to be non-zero.

measure defined in terms of occurrence mapping is a metric. In some cases where each vertex is uniquely labeled, *graph distance measure based on occurrence mapping is equivalent to that based on pattern mapping,* due to the fact that in these cases isomorphic relation is equivalent to equal relation between graphs.

In the existing structure-based graph distance metrics, only the node and edge information of a graph is used to evaluate graph distance. In other words, only occurrence mappings Γ_0 and Γ_1 are employed. Thus, from the viewpoint of occurrence mapping, existing structured-based graph distance metrics can be considered as special cases of occurrence-based graph distance metrics. Therefore, the graph distance measures defined in Definition 4.1 and 4.2 are generalization of existing structure graph distance metrics. It is just in this sense we call them *unified structure-based graph distance measures*.

As it will be shown in the experimental part of the paper, the structure difference between different graphs can be amplified when suitable Γ_i is selected, and in general, Γ_0 and Γ_1 can't capture the obvious structure difference between graphs. Hence, in real applications, rational selection of Γ_i can make the evaluation of graph distance more accurate. . Compared to the existing structure-based graph distance measures, the graph distance measures . based on substructure abundance can evaluate the graph distance in much finer grain.

4.3 Variants of Graph Distance Measures Based on Substructure Abundance

For any graph, $\Gamma_i(G)$ only captures information of those substructures with i edges. However, in some cases, the size of substructures varying in a range rather than being a fixed value will characterize the graphs better. Thus, the elementary graph distance measures that are based on substructure abundance need to be extended to include substructures with different sizes. For this purpose, it's necessary to extend graph distance defined in Definition 4.1 and 4.2 from Γ_i to Γ_I, which can capture more substructure information of a given graph. For this purpose, we will first introduce Corollary 4.3, which is an extension of Theorem 4.1 and 4.2. Then based on Corollary 4.3, we provide two variants of the substructure abundance-based graph distance measures.

Before the discussion of this section, we first give some essential notations. Let $U=\{0,1,\ldots,m\}$, where $m=|E(G)|$. Let $I\subseteq U$ and $\Gamma_I=\cup_{(i\in I)}\Gamma_i$ *s.t.* $\Gamma_I(G)=\cup_{(i\in I)}\Gamma_i(G)$, where Γ_i is a substructure mapping(a pattern mapping or an occurrence mapping). Obviously, it follows that for any integer pair (i,j) *s.t.* $i\neq j$, $\Gamma_i(G)\cap\Gamma_j(G)=\varnothing$.

Corollary 4.3: Given substructure mapping Γ_I that get all substructures(patterns or occurrences) with $i\in I$ edges. For any two labeled graphs G and G', if $G\leqq G'$, then the following statements hold:

(1)There exists an *injective mapping* ϕ: $\Gamma_I(G)\rightarrow\Gamma_I(G')$ such that for each $g\in\Gamma_I(G)$, there is only one unique $\phi(g)\in\Gamma_I(G')$ *s.t.* $g\cong\phi(g)$.

(2) $|\Gamma_I(G)|\leq|\Gamma_I(G')|$.

(3) If Γ_I is a pattern mapping, then it follows that $G\cong G'\Rightarrow|\Gamma_I(G)|=|\Gamma_I(G')|$.

If Γ_I is an occurrence mapping, then it follows that $G\cong G'\Leftrightarrow|\Gamma_I(G)|=|\Gamma_I(G')|$.

Definition 4.3: Given a substructure mapping Γ_I (.a pattern mapping or an occurrence mapping), the distance of two non-empty graphs G_1 and G_2 is defined as $d_I(G_1,G_2)=1-|\Gamma_I(G_{12})|/M(|\Gamma_I(G_1)|,|\Gamma_I(G_2)|)$, where $M(|\Gamma_I(G_1)|,|\Gamma_I(G_2)|)$ can be defined in three cases as above**.**

Theorem 4.5: For any graphs G_1,G_2 and G_3, the following properties hold true for graph distance measure defined in Definition 4.3, (1)Non-negativity, (2) Uniqueness (only Reflexivity when Γ_I is a pattern mapping), (3)Symmetry,(4)Triangle Inequality.

*Proof***:** We only prove the theorem when Γ_I is an occurrence mapping. When Γ_I is a pattern mapping, it is unnecessary to show that $d(G_1,G_2)=0\Rightarrow G_1\cong G_2$, and proofs of other properties are the same as corresponding proofs for occurrence mapping.

Since $\Gamma_i(G)\cap\Gamma_j(G)=\varnothing$ $(i\neq j)$. We have the following transformation holds.

$1\text{-}|\Gamma_I(G_{12})|/M(|\Gamma_I(G_1)|, |\Gamma_I(G_2)|) = 1\text{-}|(\Gamma_{i1}\cup\ldots\cup\Gamma_{ik})(G_{12})|/M(|(\Gamma_{i1}\cup\ldots\cup\Gamma_{ik})(G_1)|, |(\Gamma_{i1}\cup\ldots\cup\Gamma_{ik})(G_2)|)$

$=1\text{-}|(\Gamma_{i1}(G_{12})\cup\ldots\cup\Gamma_{ik}(G_{12}))|/M(|(\Gamma_{i1}(G_1)\cup\ldots\cup\Gamma_{ik}(G_1))|,|(\Gamma_{i1}(G_2)\cup\ldots\cup\Gamma_{ik}(G_2))|)$

$=1\text{-}|\Gamma_{i1}(G_{12})|+\ldots+|\Gamma_{ik}(G_{12}))|/M(|\Gamma_{i1}(G_1)|+\ldots+|\Gamma_{ik}(G_1))|,|\Gamma_{i1}(G_2)|+\ldots+|\Gamma_{ik}(G_2)|)$

$=1\text{-}\sum_{i\in I}|\Gamma_i(G_{12})|/M(\sum_{i\in I}|\Gamma_i(G_1)|, \sum_{i\in I}|\Gamma_i(G_2)|)$

(1) Non-negativity.

From Theorem 4.1, it follows that for each i, $|\Gamma_i(G_{12})| \le |\Gamma_i(G_1)|$ and $|\Gamma_i(G_{12})| \le |\Gamma_i(G_2)|$, which implies that $\sum_{i\in I}|\Gamma_i(G_{12})| \le \sum_{i\in I}|\Gamma_i(G_1)|$, and $\sum_{i\in I}|\Gamma_i(G_{12})| \le \sum_{i\in I}|\Gamma_i(G_2)|$ (eq1). Hence, we have $\sum_{i\in I}|\Gamma_i(G_{12})| \le max(\sum_{i\in I}|\Gamma_i(G_1)|, \sum_{i\in I}|\Gamma_i(G_2)|)$ (eq2).

(2) Uniqueness.

First we prove '⇒'. $d(G_1,G_2)=0 \Rightarrow \sum_{i\in I}|\Gamma_i(G_{12})| = max(\sum_{i\in I}|\Gamma_i(G_1)|, \sum_{i\in I}|\Gamma_i(G_2)|)$. Since for each i, $|\Gamma_i(G_{12})| \le |\Gamma_i(G_1)|$ and $|\Gamma_i(G_{12})| \le |\Gamma_i(G_2)|$, we have for each i, $|\Gamma_i(G_{12})|=|\Gamma_i(G_1)|=|\Gamma_i(G_2)|$. So we have $G_{12} \cong G_1 \cong G_2$.

Then we prove '⇐'. If $G_1 \cong G_2$, then for each i, we have $|\Gamma_i(G_{12})|=|\Gamma_i(G_1)|=|\Gamma_i(G_2)|$. Thus $\sum_{i\in I}|\Gamma_i(G_{12})| = max(\sum_{i\in I}|\Gamma_i(G_1)|, \sum_{i\in I}|\Gamma_i(G_2)|)$, so we have $d(G_1,G_2)=0$.

(3) Symmetry. It follows directly from the symmetry of the equation as defined in the theorem.

(4) Triangle inequality. The detailed proof of triangle inequality is shown in Appendix B. □

Definition 4.4. Given a substructure mapping Γ_I (a pattern mapping or an occurrence mapping), the distance of two non-empty graphs G_1 and G_2 is defined as $d(G_1,G_2)= \sum_{i\in I}\alpha_i d_i(G_1,G_2)$, where $\alpha_i \ge 0$ and $\sum_{i\in I}\alpha_i =1$ and $d_i(G_1,G_2)$ is a graph distance measure defined in Definition 4.1 or Definition 4.2.

Theorem 4.6. The following properties hold true for graph distance measure defined in Definition 4.4 (1)Non-negativity, (2) Uniqueness(only Reflexivity when Γ_I is a pattern mapping), (3)Symmetry,(4)Triangle Inequality.

Proof:

(1) Non-negativity. $d_i(G_1,G_2)\ge 0 \Rightarrow \alpha_i d_i(G_1,G_2) \ge 0 \Rightarrow \sum_{i\in I}\alpha_i d_i(G_1,G_2)\ge 0$

(2) Uniqueness.

First we prove '⇒'. $\sum_{i\in I}\alpha_i d_i(G_1,G_2)=0$ and $\alpha_i \ge 0$, $\sum_{i\in I}\alpha_i =1$ and $d_i(G_1,G_2) \ge 0 \Rightarrow d_i(G_1,G_2)=0$ for each $i\in I \Rightarrow G_1 \cong G_2$.

Then we prove '⇐'. $G_1 \cong G_2 \Rightarrow d_i(G_1,G_2)=0$ for each $i\in I \Rightarrow \sum_{i\in I}\alpha_i d_i(G_1,G_2)=0$.

(3) Symmetry. It follows directly from the symmetry of the equation as defined in the theorem.

(4) Triangle inequality. Triangle inequality holds true for $d_i(G_1,G_2)$ ⇒ for each $i\in I$, $d_i(G_1,G_2)+ d_i(G_2,G_3) \ge d_i(G_1,G_3) \Rightarrow$ for each $i\in I, \alpha_i d_i(G_1,G_2)+ \alpha_i d_i(G_2,G_3) \ge \alpha_i d_i(G_1,G_3) \Rightarrow \sum_{i\in I}\alpha_i d_i(G_1,G_2)+ \sum_{i\in I}\alpha_i d_i(G_2,G_3) \ge \sum_{i\in I}\alpha_i d_i(G_1,G_3)$. □

An immediate consequence of Theorem 4.6 is the following corollary.

Corollary 4.4: The following properties hold true for graph distance measure defined as $d(G_1,G_2)=(\sum_{i\in I} d_i(G_1,G_2))/k$, $(k=|\Gamma_I|)$, (1)Non-negativity, (2) Uniqueness (only Reflexivity when Γ_I is a pattern mapping), (3)Symmetry, (4)Triangle Inequality.

4.4 Variants of Unified Graph Distance Measures in Real Applications

When applying the above graph distance measures to real problems, we need to address two key issues. The first one is subgraph enumeration. The second oneis how to reasonably weight the substructure of each dimension in SVA of a graph.

To enumerate all the non-isomorphic or non-equal subgraphs of a graph is non-trivial due to the exponential growth of number of subgraphs with the increase of the size of the subgraph. However, in real world applications, it is usually not necessary to evaluate graph distance with such high precision. It is unnecessary to enumerate subgraphs with large size. Hence, the rational way to solve this problem is to customize Γ_I according to the requirements of the real applications, considering the tradeoff between the accuracy of the distance measure and computational complexity.

Since enumerating all subgraphs with i edges is time-consuming for larger i, we can restrict $\Gamma_i(G)$ to be a subset of substructures with i edges. Compared to trees and graphs, path is more simple and its enumeration is less time-consuming. Hence, we can construct substructure mappings $P=\{P_i|0\leqslant i\leqslant|E(G)|\}$ with each P_i geting all the non-isomorphic or non-equal paths with length i. Furthermore, for certain precision, it is also unnecessary to enumerate longer paths. And we will show that the graph distance measures defined according to P also possess most properties of a metric.

Corollary 4.5. $\vec{V}=(|P_0|,\ldots,|P_m|)$, $m=|E(G)|$ is a graph invariant.

Corollary 4.6. Let $U=\{0,1,\ldots,m\}$,$m=|E(G_{12})|$, $I\subseteq U$, then graph distance measure $d(G_1,G_2)=1-|P_I(G_{12})|/M(|P_I(G_1)|,|P_I(G_2)|)$ with $G_{12}=mces(G_1,G_2)$ and $M(|P_I(G_1)|,|P_I(G_2)|)$ defined in three cases as before, satisfies the following properties (1)Non-negativity, (2) Uniqueness (only Partial Uniqueness when P_I is a pattern mapping), (3)Symmetry,(4)Triangle Inequality.

Corollary 4.7. The following properties hold true for graph distance measure defined as $d(G_1,G_2)=(\sum_{i\in I} d_i(G_1,G_2))/k$, ($k=|P_I|$), (1)Non-negativity, (2) Uniqueness (only Reflexivity when P_I is a pattern mapping), (3)Symmetry,(4)Triangle Inequality.

To address the second issue, we must be aware that different substructures of a graph can not characterize the graph to the same extent. And a basic observation is that two graphs are more similar to each other if they share more *complex and unique* substructures instead of simple and trivial structures such as isolated nodes or edges. Hence, different subgraphs appearing in a common graph of G_1 and G_2 will have different contribution to the similarity of these two graphs, and the occurrence of complex and unique substructures in the common graph will be a significant indication of similarity between graphs.

Thus, we need to give the definition of the uniqueness of a subgraph. Informally, similar to the uniqueness used in [9, 15], the *uniqueness* of a subgraph $g\subseteq G$ can be evaluated according to the frequence of its occurrence in random graphs with size equivalent to G. Let $f_{rand}(g)$ be the frequency of occurrence g in a randomized network G_{randi}, for $1\leqslant i\leqslant N$, where N is the number of randomized networks and each randomized network has $|V(G)|$ nodes, and nodes are linked by probability $p=2|E(G)|/(|V(G)|*(|V(G)|-1))$. Then the uniqueness of subgraph g can be described by $uniq(g,G)=1-f_{rand}(g)/N$.

In the definition of graph distance measure, we can assign to each dimensional substructure a weight, which is computed according to the uniqueness of substructures of the graph. For example, if the graph distance is defined according to $\Gamma'\subseteq\Gamma$, then for each $\Gamma_i\in\Gamma'$, we can get an average uniqueness $avg(\Gamma_i)=(\sum_{g\in\Gamma i(G)}uniq(g,G))/|\Gamma_i(G)|$. Furthermore, we would normalize $avg(\Gamma_i)$ and let $\nabla avg(\Gamma_i)=avg(\Gamma_i)/\sum avg(\Gamma_i)$. Obviously, $\nabla avg(\Gamma_i)\geq 0$ and $\sum\nabla avg(\Gamma_i)=1$. Hence, it's not difficult to get the following corollary.

Corollary 4.8. Given a substructure mapping Γ_I (a pattern mapping or an occurrence mapping), the following properties hold true for graph distance measure defined as $d(G_1,G_2)=\sum_{i\in I}\alpha_i d_i(G_1,G_2)$, where $\alpha_i=\nabla avg(\Gamma_i)$. , (1)Non-negativity, (2)Uniqueness (only Reflexivity when P_I is a pattern mapping), (3) Symmetry, (4) Triangle Inequality.

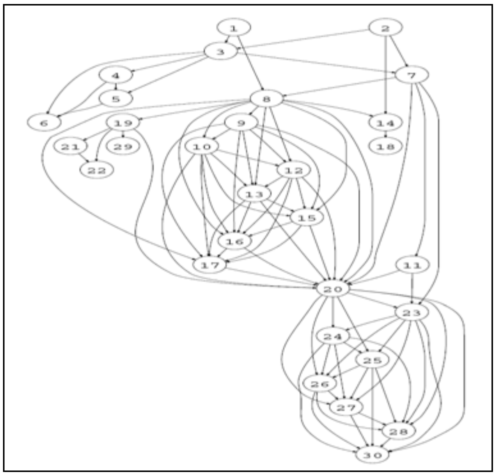


Figure 5: Population Bayesian Network of HAN

5 Application in Population Structure analysis

In this section, we will apply the graph distance measures defined in the previous sections to population structure analysis. We will demonstrate the precision of these graph distance measures through this example.

5.1 Bayesian Marker Networks for Three Populations

With the accomplishment of Human Genome Project and International HapMap Project [20], large amounts of sequences and genotype data are available and they provide good sources for the population structure study. Typed SNPs(Single nucleotide polymorphisms)[8] can be used to construct Bayesian marker network that models the dependence relations (linkage disequilibrium) among markers [10]. Due to evolutions, linkage disequilibrium between the markers varies across populations. The differences in the structure of Bayesian networks between populations imply the different history of population evolution. Therefore, the distance between the marker networks will correspond the distance between populations. To evaluate the performance of the proposed graph distance measures, we typed 30 SNPs from the Chromosome 21 for 48 individuals from African American population (AFA), 46 individuals from Chinese Han Population (HAN), and 40 individuals from European Caucasian population (CAU), to create three Bayesian marker networks for three populations. We use directed graph to represent the Bayesian networks, a node in the graph denotes a SNP marker. The mutual information between two markers is calculated, which approximately measures the linkage disequilibrium between two markers [11]. The constructed Bayesian network of HAN with 30 SNPs is shown in Figure 5. The other two networks are close to this one, thus not shown below. The numbers of edges of Bayesian marker networks for AFA, HAN and CAU populations are 89, 90,116, respectively. The average node degrees of three networks are 2.97, 3.00, and 3.87, respectively.

5.2 Population Structure Analysis

The graph distance measures are applied to measuring the distance between populations. We enumerate all the simple paths of three networks. The path length distributions of three marker networks are shown in Figure 6. From the figure, we can see that AFA contains the least number of paths, while HAN contains the longest paths. And it is clear that the difference of substructure abundance between these three networks is much more obvious when regarding to the middle size of substructures. Hence, it is rational to measure the graph distance with respect to the middle-size substructures of the graphs.

Figure 7 shows the path number distributions of the maximum common edge-induced subgraph of three pairs of networks. The path number of common graphs represents the absolute similarity between populations. Figure 8 shows the relative similarity between three populations, which is the ratio of common path number to the problem of size. In this experiment, we use $|P_i(G_1)|+|P_i(G_2)|-|P_i(G_{12})|$ to measure the size of problem. Note that for both relative and absolute distance, we can not discern the difference among three population pairs when path length is very small or very large.

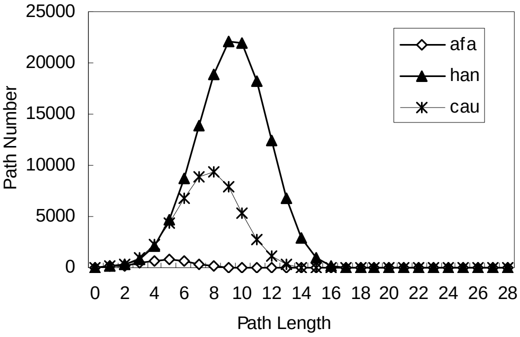


Figure 6: Path Distribution of Three Populations

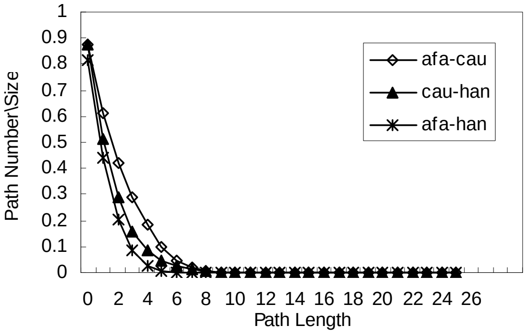


Figure 8: Relative Similarities among populations

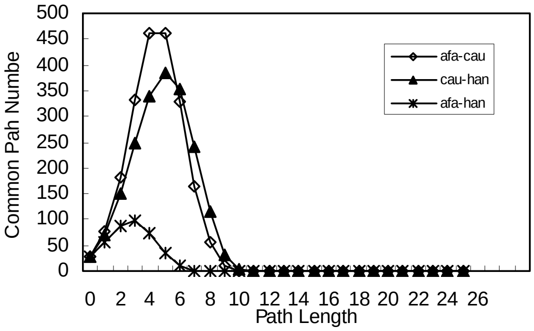


Figure 7: Absolute Similarities among populations

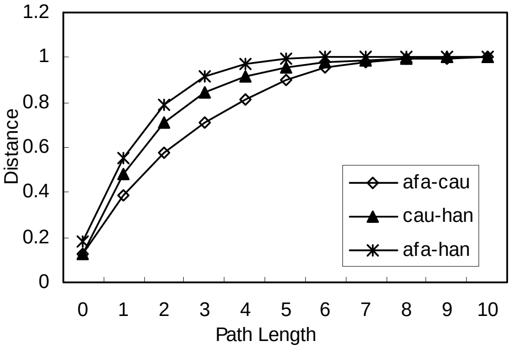


Figure 9: Graph Distance among populations for each P_i

For $0 \le i \le 10$, we work out the graph distance of each pair of three populations according to graph distance measures $d(G_1,G_2)=1-|P_i\ (mces(G_1,\ G_2))|/max(|P_i\ (G_1)|,|P_i\ (G_2)|)$ for each P_i. The graph distances among populations for different path lengths are shown in Table 1 and corresponding plot is shown in Figure 9. For $i>10$, each common graph contains no substructure of size i, thus graph distance measured with respect to corresponding P_i is trivial.

We also calculate graph distances according to the distance measures with respect to P_I. The result is shown in Table 2. We use 'sum[i, j]' to denote the graph distance measure whose similarity is defined by the cardinality of P_I (G) with $I=[i, j]$, *i.e.* the graph distance measure defined in Corollary 4.6. We use 'avg[i, j]' to denote the average graph distance over P_I with $I=[i, j]$, *i.e.* the graph distance measure defined in Corollary 4.7. We compute 'avg' and 'sum' in the range [2, 3], because from Figure 8 we can see that P_2 and P_3 can capture the most obvious substructure difference among three population networks. We also compute 'sum[0,1]', which is another usually used graph distance measures in many real applications.

Table 1: Graph distances among populations for each P_i

	Path Length										
Population	0	1	2	3	4	5	6	7	8	9	10
AFA-CAU	0.125	0.386	0.580	0.709	0.817	0.900	0.953	0.982	0.994	0.999	1.000
CAU-HAN	0.125	0.485	0.708	0.843	0.917	0.956	0.977	0.989	0.996	0.999	1.000
AFA-HAN	0.182	0.556	0.793	0.916	0.972	0.994	0.999	1.000	1	1	1

Table 2: Graph distances among populations of P_I

	Path Length Range		
Population	Sum[0,1]	Avg[2,3]	Sum[2,3]
AFA-CAU	0.333	0.644	0.674
CAU-HAN	0.417	0.776	0.811
AFA-HAN	0.478	0.855	0.884

Table 3: Graph distances with distance between AFA-HAN normalized as 1

	Path Length Range				
Population	Node	Edge	Sum[0,1]	Avg[2,3]	Sum[2,3]
AFA-CAU	0.687	0.694	0.697	0.753	0.762
CAU-HAN	0.687	0.872	0.872	0.907	0.917
AFA-HAN	1	1	1	1	1

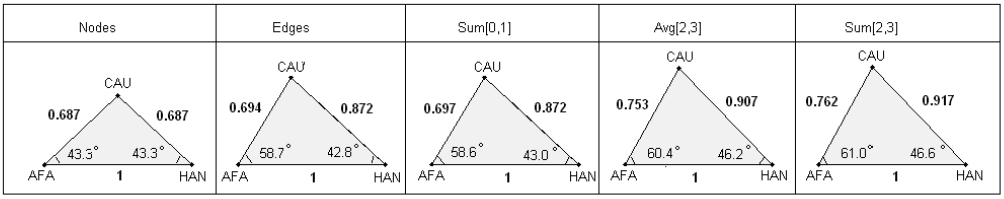


Figure 10: Distance graph of population structures under different distance measures

At last, we draw out five distance graphs among these three populations for graph distance measure defined according to P_0, P_1 , *sum*($P_{[0,1]}$), *avg*($P_{[2,3]}$) and *sum*($P_{[2,3]}$), respectively. For the convenience of observation, we normalize the distance value between AFA and HAN to 1. The normalized detailed distance values are shown in Table 3 and the corresponding distance graphs are drawn in Figure 10.

Among all these graph distance measures, we believe that 'sum[2,3]' is the most appropriate graph distance in this case, which can amplify the minute distance difference. In population structure analysis, this kind of minute difference can lead to the wrong qualitative assertion. For instance, if only Γ_0 is used in the measurement of the graph distance, we can conclude that the distance between CAU and AFA is the same as it between CAU and HAN. However, when 'sum[2, 3]' is employed, it is clear that CAU is much closer to AFA than HAN.

The results show that the distances between HAN and other two populations are the furthest.while the distance between CAU and AFA is shorter, which implicates the SNPs linkage disequilibrium structure of HAN population is more complex.

6 Related Works

Structure-based graph distance measures have been widely studied in pattern recognition and chemical informatics area. Bunke and Shearer [1] first proposed graph distance metric based on maximal common graph, which underlies following structure-based graph distance measures. In their pioneering works, $|max(|G_1|,|G_2|)|$ is used as the problem size, which ignores the influence of the smaller one of the two graphs. Bunke[2,3] also revealed the relation between MCS-based graph distance and graph edit distance, which bridges the structure-based graph distance and traditional graph edit distances that are based on cost functions.

Hereafter, a variety of structure-based distance metrics have been proposed. Wallis et al [19] proposed graph distance based on graph union, where $|G_1|+|G_2|-|G_{12}|$ is employed as the size of the problem. Then, Fernandez and Valiente [5] evaluated the distance between graphs by measuring the missing structural information expressed as the difference between minimal common supergraph and maximal common subgraph. Dzena Hidovic and Marcello Pelillo [6] developed two attributed graph distance metrics based on the precedent structured graph distance metric framework.

All the above graph distance metrics except [6] have been systematically surveyed by John W. Raymond and Peter Willett [14]. A series of John W. Raymond 's works [12,13,14,15] have focused on virtual screening through evaluating the distance of chemical compounds. The most important contribution of John W. Raymond 's work is RASCAL[12], an efficient graph similarity calculation procedure, in which many efficient similarity filtering strategies have been employed and an efficient maximum common subgraph isomorphism detection algorithm has been devised.

Bayesian Network is an abstract presentation of complex networks, which provide a new tool for studies of the structure of biological system. Many approaches based on Bayesian methods to study the gene regulation and protein-protein interaction network are brought forward [7, 16, 17]. However, these studies focused on the functional perspective, and the structure study of the sequences which constitute gene and translate to protein is very little. SNPs are common single base variation in the human genome sequence. They play an important role in the association analysis of complex diseases. The complexities of SNPs linkage disequilibrium are important features of population evolution. Constructing Bayesian network with SNPs from different populations is meaningful for the studies of population evolution. It turn out that Bayesian networks of SNPs will open a new field in the network approach to studies of population structure and evolution.

7 Conclusion

In this paper, to evaluate graph distance in high degree of precision, we proposed unified structure-based graph distance measures and their variants, utilizing substructure abundance vector. We employ these graph distance measures to calculate the distances between populations in population structure analysis, where accurate evaluation of graph distance is desired.

In future ., it is of great interest to study the relation between substructure abundance and the symmetry of a graph so that more theoretic algebraic tools can be used to perform deeper research on graph distance measure theory. Another significant work is to use the graph distance measures proposed in this paper to construct the distance graph of more population structures, which will unravel more accurate population structures of the genetic data. The results in this paper are very limited, we plan to perform large-scale calculations of the graph distance measures proposed in this paper in more real applications.

Acknowledgments

This research was supported by the National Natural Science Foundation of China under Grant No.60303008, No.60673133, No.60703093; the National Grand Fundamental Research 973 Program of China under Grant No. 2005CB321905.

References

[1] H. Bunke and K. Shearer. A graph distance metric based on the maximal common subgraph. Pattern Recognition Letters, 19:255–259, 1998.

[2] H. Bunke. Error correcting graph matching: On the influence of the underlying cost function. IEEE Transactions on Pattern Analysis and Machine Intelligence, 21(9):917–922, 1999.

[3] H. Bunke. On a Relation between Graph Edit Distance and Maximum Common Subgraph, Pattern Recognition Letters, vol. 18, pp. 689-694, 1997.

[4] J. Chen, W. Hsu, ML. Lee, and SK. Ng. NemoFinder:Dissecting genome-wide protein-protein interactions with meso-scale network motifs. KDD, 106-115,2006.

[5] M.L.Fernandez and G.Valiente. A graph distance metric combining maximum common subgraph and minimum common supergraph. Pattern Recognition Letters, 22:753–758, 2001.

[6] Dzena Hidovic, Marcello Pelillo: Metrics For Attributed Graphs Based On The Maximal Similarity Common Subgraph. IJPRAI 18(3): 299-313 (2004)

[7] Ronald Jansen, Haiyuan Yu, Mark Gerstein, et al., A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data. Science,2003. 302: 449-453

[8] L. Kruglyak, and D.A. Nickerson, Variation is the spice of life. Nat Genet, 2001. 27(3): 234-6

[9] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. Science, 298:824–827, 2002.

[10] K P Murphy. A Brief Introduction to Graphical Models and Bayesian Networks [Z].2001.

[11] M Nothnagel, R Furst, K Rohde. Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks[J]. Hum Hered, 2002, 54:186-198.

[12] John W. Raymond, Eleanor J. Gardiner, Peter Willett: RASCAL: Calculation of Graph Similarity using Maximum Common Edge Subgraphs. Comput. J. 45(6): 631-644 (2002)

[13] John W. Raymond, Eleanor J. Gardiner, Peter Willett: Heuristics for Similarity Searching of Chemical Graphs Using a Maximum Common Edge Subgraph Algorithm. Journal of Chemical Information and Computer Sciences 42(2): 305-316 (2002)

[14] John W. Raymond, Peter Willett: Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2D chemical structure databases. Journal of Computer-Aided Molecular Design 16(1): 59-71 (2002)

[15] John W. Raymond, Peter Willett: Maximum common subgraph isomorphism algorithms for the matching of chemical structures. Journal of Computer-Aided Molecular Design 16(7): 521-533 (2002)

[16] Daniel Rhodes, Arul M Chinnaiyan, et al., Probabilistic model of the human protein-protein interaction network. Nature Biotechnology, 2005. 23(8):951-959

[17] E Segal, M Shapira, A Regev, et al., Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nature Genetics, 2003. 34(2):166-176.

[18] M. Stephens, N.J. Smith, and P. Donnelly, A new statistical method for haplotype reconstruction from population data. Am J Hum Genet, 2001. 68(4):978-89.

[19] W.D. Wallis, P. Shoubridge, M. Kraetz, and D. Ray. Graph distances using graph union. Pattern Recognition Letters, 22:701–704, 2001.

[20] The International HapMap Consortium, The International HapMap Project. Nature, 2003. 426:789-796.

[21] B.T. Messmer, H. Bunke, A new algorithm for error-tolerant subgraph isomorphism detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (5) (1998) 493–504.

[22] Michel Neuhaus and Horst Bunke ,Automatic learning of cost functions for graph edit distance, Information Sciences, Volume 177, Issue 1, 2007, Pages 239-247.

[23] Xifeng Yan; Feida Zhu; Yu Philip S.; Jiawei Han, Feature-based similarity search in graph structures ACM Transaction On Database Systems,. 31(4):1418–1453, 2006.

[24] Matthias Dehmer and Frank Emmert-Streib, Structural similarity of directed universal hierarchical graphs: A low computational complexity approach, Applied Mathematics and Computation, Volume 194, Issue 1, 2007, Pages 7-20.

[25] F. Chevalier, J.-P. Domenger, J. Benois-Pineau and M. Delest , Retrieval of objects in video by similarity based on graph matching Pattern Recognition Letters,Volume 28, Issue 8, 2007, Pages 939-949.

[26] Matthias Dehmer and Frank Emmert-Streib, Comparing large graphs efficiently by margins of feature vectors, Applied Mathematics and Computation, Volume 188, Issue 2, 2007, Pages 1699-1710.

[27] Matthias Dehmer, Frank Emmert-Streib and Jürgen Kilian ,A similarity measure for graphs with low computational complexity Applied Mathematics and Computation, Volume 182, Issue 1, 2006, Pages 447-459.

[28] Sergio Flesca, Giuseppe Manco, Elio Masciari, Luigi Pontieri and Andrea Pugliese, Exploiting structural similarity for effective Web information extraction. Data & Knowledge Engineering, Volume 60, Issue 1, 2007, Pages 222-234.

Appendix A:

Theorem A.1. Let Γ_i be a substructure mapping (pattern mapping or occurrence mapping), for any three graphs G_1, G_2 and G_3 , *triangle inequality* holds true for graph distance measure $d_i(G_1,G_2)=1-|\Gamma_i(G_{12})|/max(|\Gamma_i(G_1)|,|\Gamma_i(G_2)|)$.

Proof: Suppose we have three graphs denoted by G_1, G_2, G_3. For the notational convenience, let $m_1=|\Gamma_i(G_1)|$, $m_2=|\Gamma_i(G_2)|$, $m_3=|\Gamma_i(G_3)|$, $m_{12}=|\Gamma_i(G_{12})|$, $m_{23}=|\Gamma_i(G_{23})|$, $m_{13}=|\Gamma_i(G_{13})|$. Then we have
$d_i(G_1,G_2)=1-m_{12}/max(m_1, m_2)$,
$d_i(G_2,G_3)=1-m_{23}/max(m_2, m_3)$,
$d_i(G_1,G_3)=1-m_{13}/max(m_1, m_3)$.
To prove the triangle inequality holds true for the graph distance measure equals to show:
$d_i(G_1,G_2)+ d_i(G_2,G_3) \geq d_i(G_1,G_3)$, *i.e.*
$(1-m_{12}/max(m_1, m_2))+(1-m_{23}/max(m_2, m_3))\geq 1-m_{13}/max(m_1, m_3)$ (*)

There are six possible cases need to be distinguished and proven.

Case 1: $G_{12}=G_{23}=G_{13}=\varnothing$. (Notice that if the graphs are unlabeled, this case will never happen.)
This means $m_{12}=m_{23}=m_{13}=0$. So (*) can be reduced to 1+1≥1, which is trivial.

Case 2: Only one of G_{12}, G_{23}, G_{13} is non-empty
(1)Suppose only $G_{12}\neq\varnothing$, then $m_{23}= m_{13}=0$, (*) will be reduced to
$2-m_{12}/ max(m_1,m_2)\geq 1$ *i.e.* $1\geq m_{12}/max(m_1,m_2)$
Since $G_{12}\leqq G_1$ and $G_{12}\leqq G_2$, then $m_{12}\leq m_1$ and $m_{12}\leq m_2$ according to Theorem 4.1 and 4.2. Thus $m_{12} \leq max(m_1,m_2)$ and the above inequality holds.
(2)Suppose only $G_{23}\neq\varnothing$, then $m_{12}=m_{13}=0$, (*) will be reduced to
$2-m_{23}/max(m_2,m_3)\geq 1$, the following proof process is the same as (1).
(3)Suppose only $G_{13}\neq\varnothing$, then $m_{12}=m_{23}=0$, (*) will be reduced to
$2\geq 1-m_{13}/max(m_1,m_3)$, *i.e* $m_{13}/max(m_1, m_3)\geq -1$, which is trivial.

Case 3: Only one of G_{12}, G_{23}, and G_{13} is empty.
(1)Suppose only $G_{13}=\varnothing$, then $m_{13}=0$, inequality (*) will be reduced to
$1-m_{12}/max(m_1,m_2)-m_{23}/max(m_2, m_3)\geq 0$ (3.1)
a) $m_1\geq m_2\geq m_3$, then (3.1) is equivalent to $1-m_{12}/m_1 -m_{23}/m_2\geq 0$, *i.e.* $m_1m_2 - m_2m_{12} - m_1m_{23}\geq 0$.
Since $m_1\geq m_2$, $m_1m_2 - m_2m_{12} - m_1m_{23}\geq m_1m_2 - m_1m_{12} - m_1m_{23}$, *i.e.*
$m_1m_2 - m_2m_{12} - m_1m_{23}\geq m_1(m_2 -m_{12} -m_{23})$ (3.2)

Since $G_{13}=\varnothing$, G_1 and G_3 have no common subgraphs. This implies that $\forall g\in\Gamma_i(G_{12})$ and $\forall g'\in\Gamma_i(G_{23})$, g is not isomorphic to g'. Obviously, we have $G_{12}\leqq G_2$ and $G_{23}\leqq G_2$. According to Corollary 4.1, we have $m_{12}+m_{23}\leq m_2$, which shows (3.2) ≥0. Thus we can prove that (3.1) holds.
b) $m_1\geq m_3\geq m_2$, then (3.1) is equivalent to $1-m_{12}/m_1-m_{23}/m_3\geq 0$, *i.e.* $m_1m_3 - m_3m_{12} - m_1m_{23}\geq 0$.
Since $m_1\geq m_3$, $m_1m_3 - m_3m_{12} - m_1m_{23}\geq m_1m_3 - m_1m_{12} - m_1m_{23}$
$\geq m_1m_2 - m_1m_{12} - m_1m_{23}$ (due to $m_3\geq m_2$), *i.e.* $m_1m_3 - m_1m_{12} - m_1m_{23}\geq m_1(m_2 -m_{12} -m_{23})$
Notice that this is exactly the inequality (3.2), so the following proof is the same.
c) $m_2\geq m_1\geq m_3$, then (3.1) is equivalent to $1-m_{12}/m_2 -m_{23}/m_2\geq 0$,
i.e. $m_2 - m_{12} - m_{23} \geq 0$, which we have proved in a).
d) $m_2\geq m_3\geq m_1$, the proof is the same as c).

e) $m_3 \geq m_1 \geq m_2$, then (3.1) is equivalent to $1\text{-}m_{12}/m_1\ \text{-}m_{23}/m_3 \geq 0$, *i.e.* $m_1m_3 - m_3m_{12} - m_1m_{23} \geq 0$.

Since $m_1m_3 - m_3m_{12} - m_1m_{23} \geq m_1m_3 - m_3m_{12} - m_3m_{23} = m_3(m_1 - m_{12} - m_{23}) \geq m_3 * (m_2 - m_{12} - m_{23})$, which is the same as inequality(3.2), so the following proof is the same as a).

f) $m_3 \geq m_2 \geq m_1$, then (3.1) is equivalent to $1\text{-}m_{12}/m_2\ \text{-}m_{23}/m_3 \geq 0$,*i.e.* $m_2m_3 - m_3m_{12} - m_2m_{23} \geq 0$.

Since $m_2m_3 - m_3m_{12} - m_2m_{23} \geq m_2m_3 - m_3m_{12} - m_3m_{23} = m_3(m_2 - m_{12} - m_{23})$, which is the same as inequality(3.2), so the following proof is the same.

Problem: can we proof directly like this? To prove:

$1\text{-}m_{12}/max(m_1,m_2)\text{-}m_{23}/max(m_2, m_3) \geq 0$ (3.1)

$max(m_1,m_2) >= m_2\ max(m_2, m_3)\ ) >= m_2 \quad m_{12}/max(m_1,m_2)+m_{23}/max(m_2, m_3) <= (\ m_{12}/m_2 + m_{23}/m_2)$

Since $G_{13}=\varnothing$, G_1 and G_3 have no common subgraphs. This implies that $\forall g \in \Gamma_i(G_{12})$ and $\forall g' \in \Gamma_i(G_{23})$, g is not isomorphic to g'. Obviously, we have $G_{12} \leqq G_2$ and $G_{23} \leqq G_2$. According to Corollary 4.1, we have $m_{12}+m_{23} \leq m_2$, which shows (3.2) ≥0. Thus we can prove that (3.1) holds.

(2)Suppose only $G_{12}=\varnothing$, then $m_{12}=0$, (*) will be reduced to

$2\text{-}m_{23}/max(m_2 ,m_3) \geq 1\text{-}m_{13} / max(m_1 ,m_3)$.

Since $m_{23}/max(m_2 ,m_3) \leq 1$, $2\text{-}m_{23} / max(m_2 ,m_3) \geq 1 \geq 1\text{-}m_{13} / max(m_1 ,m_3)$.

(3) Suppose only $G_{23}=\varnothing$, then $m_{23} =0$, (*) will be reduced to

$2\text{-}m_{12}/max(m_1 ,m_2) \geq 1\text{-}m_{13}/max(m_1 , m_3)$.

Since $m_{12}/max(m_1 ,m_2) \leq 1$, $2\text{-}m_{12} /max(m_1 ,m_2) \geq 1 \geq 1\text{-}\ m_{13}/max(m_1 , m_3)$, which accomplishes our proof of Case 3.

Case 4: G_{12}, G_{23}, G_{13} all exist, *i.e.* $G_{12} \neq \varnothing$, $G_{23} \neq \varnothing$, and $G_{13} \neq \varnothing$.

We use G_{123} to denote the maximum common subgraph of G_1, G_2, G_3. Obviously, $G_{123} \leqq G_{12}$, $G_{123} \leqq G_{23}$ and $G_{123} \leqq G_{13}$. The overlapping between $\Gamma_i(G_1)$, $\Gamma_i(G_2)$ and $\Gamma_i(G_3)$ is shown in Figure A.1. According to Theorem 4.1 and 4.2, there is an injective mapping $\alpha:\Gamma_i(G_{123}) \rightarrow \Gamma_i(G_{12})$. Similarly, injective mapping $\beta:\Gamma_i(G_{123}) \rightarrow \Gamma_i(G_{23})$, $\gamma:\Gamma_i(G_{123}) \rightarrow \Gamma_i(G_{13})$ also exist.

Let $\Gamma_i(G'_{12})=\Gamma_i(G_{12})\text{-}\alpha^{-1}(\Gamma_i(G_{123}))$, $\Gamma_i(G'_{23})= \Gamma_i(G_{23})\text{-}\ \beta^{-1}(\Gamma_i(G_{123}))$, $\Gamma_i(G'_{13})= \Gamma_i(G_{13})\text{-}\gamma^{-1}(\Gamma_i(G_{123}))$

Let $m_{12}'=|\Gamma_i(G'_{12})|$, $m_{23}'=|\Gamma_i(G'_{23})|$ and $m_{13}'=|\Gamma_i(G'_{13})|$.

Based on this definition, it's easy to see that we have:

$m_{12} = m_{12}'+m_{123}$ (4.a)

$m_{23} = m_{23}'+ m_{123}$ (4.b)

$m_{13} = m_{13}'+ m_{123}$ (4.c)

For notational convenience, we use '$A \cap B=\varnothing$' to denote the statement that 'for two graph sets A, B, $\forall g_1 \in A$, $\forall g_2 \in B$, g_1 is not isomorphic to g_2', which can be considered as an extension of set join operating from equal to isomorphic relation between elements of a set.

Thus, we can see that the following equations (4.d)-(4.l) hold true. As an example, we will show the correctness of equation (4.d). Assume that $\Gamma_i(G_1) \cap \Gamma_i(G'_{23}) \neq \varnothing$, then there exist graphs g, $g_1 \in \Gamma_i(G_1)$ and $g_2 \in \Gamma_i(G'_{23})$, such that $g \cong g_1 \cong g_2$. Due to $\Gamma_i(G'_{23}) \subseteq \Gamma_i(G_{23})$, we have $g \leqq G_{23}$, which implies that $g \leqq G_2$ and $g \leqq G_3$. From $g \cong g_1$,$g_1 \in \Gamma_i(G_1)$, we also have $g \leqq G_1$. Thus we can conclude that $g \leqq G_{123}$, which contradict to $g \cong g2 \in \Gamma_i(G'_{23})=\Gamma_i(G_{23})\text{-}\beta^{-1}(\Gamma_i(G_{123}))$.

$\Gamma_i(G_1) \cap \Gamma_i(G'_{23})= \varnothing$ (4.d)

$\Gamma_i(G_2) \cap \Gamma_i(G'_{13})= \varnothing$ (4.e)

$\Gamma_i(G_3) \cap \Gamma_i(G'_{12})= \varnothing$ (4.f)

$\Gamma_i(G_{123}) \cap \Gamma_i(G'_{12})= \varnothing$ (4.g)

$\Gamma_i(G_{123}) \cap \Gamma_i(G'_{23})= \varnothing$ (4.h)

$\Gamma_i(G_{123})\cap\Gamma_i(G'_{13})=\varnothing$ (4.i)

$\Gamma_i(G'_{12})\cap\Gamma_i(G'_{23})=\varnothing$ (4.j)

$\Gamma_i(G'_{23})\cap\Gamma_i(G'_{13})=\varnothing$ (4.k)

$\Gamma_i(G'_{31})\cap\Gamma_i(G'_{12})=\varnothing$ (4.l)

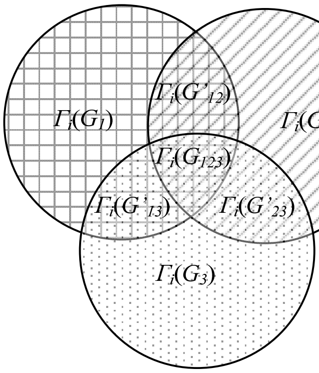


Figure A.1: Illustration of overlapping between $\Gamma_i(G_1)$, $\Gamma_i(G_2)$ and $\Gamma_i(G_3)$

There are six possible cases need to be discussed.

a) $m_1\geq m_2\geq m_3$

In this case , inequality (*) will be reduced to $(1-m_{12}/m_1)+(1-m_{23}/m_2)\geq 1-m_{13}/m_1$, *i.e.* $1-m_{12}/m_1-m_{23}/m_2+m_{13}/m_1\geq 0$, i.e. $m_1m_2-m_2m_{12}-m_1m_{23}+m_2m_{13}\geq 0$, i.e.

$m_1(m_2-m_{23})+m_2(m_{13}-m_{12})\geq 0$ (4.1)

Since $m_1\geq m_2$, $m_1(m_2-m_{23})+m_2(m_{13}-m_{12})\geq m_2(m_2-m_{23})+m_2(m_{13}-m_{12})\geq$ $m_2(m_2-m_{23}+m_{13}-m_{12})$. (4.2)

Further more, due to (4.a), (4.b) and (4.c), $m_2-m_{23}+m_{13}-m_{12}=$ $m_2-(m_{23}'+m_{123})+(m_{13}'+m_{123})-(m_{12}'+m_{123})=m_2-m_{23}'+m_{13}'-m_{12}'-m_{123}$ $=(m_2+m_{13}')-(m_{23}'+m_{12}'+m_{123})$. (4.3)

Due to (4.j), (4.g), (4.h), it follows that $\forall g_1\in\Gamma_i(G'_{12})$, $\forall g_2\in\Gamma_i(G'_{23})$ and $\forall g_3\in\Gamma_i(G_{123})$, g_1, g_2 and g_3 are pairwise non-isomorphic. Since $\Gamma_i(G'_{12})\subseteq\Gamma_i(G_{12})$, $\Gamma_i(G'_{23})\subseteq\Gamma_i(G_{23})$, according to corollary 2, we have $m_{23}'+m_{12}'+m_{123}\leq m_2$. Hence we have (4.3) ≥0 and (4.1) holds.

b) $m_1\geq m_3\geq m_2$

In this case, inequality (*) is equivalent to: $(1-m_{12}/m_1)+(1-m_{23}/m_3)\geq 1-m_{13}/m_1$, *i.e.* $1-m_{12}/m_1-m_{23}/m_3+m_{13}/m_1\geq 0$, i.e. $m_1m_3-m_3m_{12}-m_1m_{23}+m_3m_{13}\geq 0$, i.e.

$m_1(m_3-m_{23})+m_3(m_{13}-m_{12})\geq 0$ (4.4)

Since $m_1(m_3-m_{23})+m_3(m_{13}-m_{12})\geq m_3(m_3-m_{23})+m_3(m_{13}-m_{12})=m_3(m_3-m_{23}+m_{31}-m_{12})\geq m_3(m_2-m_{23}+m_{13}-m_{12})$, which is similar to (4.2) and the following proof is the same as a).

c) $m_2\geq m_1\geq m_3$

In this case, inequality (*) will be reduced to: $(1-m_{12}/m_2)+(1-m_{23}/m_2)\geq 1-m_{13}/m_1$, *i.e.* $1-m_{12}/m_2-m_{23}/m_2+m_{13}/m_1\geq 0$, *i.e.* $m_1m_2-m_1m_{12}-m_1m_{23}+m_2m_{13}\geq 0$.

Since $m_2\geq m_1$, $m_1m_2-m_1m_{12}-m_1m_{23}+m_2m_{13}\geq m_1m_2-m_1m_{12}-m_1m_{23}+m_1m_{13}=m_1(m_2-m_{23}+m_{13}-m_{12})$, which is similar to (4.2) and the following proof is the same as a).

d) $m_2\geq m_3\geq m_1$

In this case, inequality (*) will be reduced to $(1-m_{12}/m_2)+(1-m_{23}/m_2)\geq 1-m_{13}/m_3$, *i.e.* $1-m_{12}/m_2-m_{23}/m_2+m_{13}/m_2\geq 0$, *i.e.* $m_3m_2-m_3m_{12}-m_3m_{23}+m_2m_{13}\geq 0$

Since $m_2 \geq m_3$, $m_3 m_2$ - $m_3 m_{12}$ - $m_3 m_{23}$ + $m_2 m_{13} \geq$ $m_3 m_2$ - $m_3 m_{12}$ - $m_3 m_{23}$ + $m_3 m_{13} = m_3$ (m_2 - m_{23} + m_{13} - m_{12}), which is similar to (4.2) and the following proof is the same as a).

e) $m_3 \geq m_1 \geq m_2$

In this case, inequality (*) will be reduced to $(1\text{-}m_{12}/m_1)+(1\text{-}m_{23}/m_3) \geq 1\text{-}m_{13}/m_3$, *i.e.* $1\text{-}m_{12}/m_1$ $\text{-}m_{23}/m_3$ $+m_{13}/m_3 \geq 0$, *i.e.* $m_3 m_1$ - $m_3 m_{12}$ - $m_1 m_{23}$ + $m_1 m_{13} \geq 0$, i.e. $m_3(m_1$ - m_{12})+ $m_1($ m_{13} $\text{-}m_{23}$ $) \geq 0$.

Since $m_3 \geq m_1$, $m_3(m_1$ - m_{12})+ $m_1($ m_{13} $\text{-}m_{23}$ $) \geq$ $m_1(m_1$ $\text{-}m_{12}$)+ $m_1($ m_{13} $\text{-}m_{23}$ $) =$ m_1 $(m_1$ $\text{-}m_{12}$ $+m_{13}\text{-}m_{23}$)
$\geq m_1(m_2$ - m_{12} + m_{13} $\text{-}m_{23}$), which is similar to (4.2) and the following proof is the same as a).

f) $m_3 \geq m_2 \geq m_1$

In this case, inequality (*) will be reduced to $(1\text{-}m_{12}/m_2)+(1\text{-}m_{23}/m_3) \geq 1\text{-}m_{13}/m_3$, *i.e.*
$1\text{-}m_{12}/m_2$ $\text{-}m_{23}/m_3$ $+m_{13}/m_3 \geq 0$, *i.e.* $m_3 m_2$ - $m_3 m_{12}$ - $m_2 m_{23}$ + $m_2 m_{13} \geq 0$, *i.e.* $m_3(m_2$ - m_{12}) + $m_2(m_{13}$ $\text{-}m_{23}$ $) \geq 0$
Since $m_3 \geq m_2$, $m_3(m_2$ - m_{12}) + $m_2(m_{31}$ $\text{-}m_{23}$ $) \geq$ $m_2(m_2$ - m_{12}) + $m_2(m_{13}$ $\text{-}m_{23}$ $) =$ $m_2(m_2$ - m_{12} $+m_{31}$ $\text{-}m_{23}$), which is exactly the inequality (4.2) and the following proof is the same as a).

Appendix B:

In this section, we would show that the triangle inequality holds true for the graph distance measure defined in Definition 4.3. For this purpose, we will first show that arbitrary graph space can be transformed into an equivalent simplified graph space when addressing those issues related to graph isomorphism, which will be discussed in Lemma B.1 and Theorem B.1. Specifically, arbitrary graph space ***G*** as well as a graph distance measure *d* defined on itself, denoted as (***G***, *d*), can be mapped to an isomorphic graph space (***G'***, *d*), such that the quantification relation implicated by *d* is conserved. In the new graph space, triangle inequality can be easily calculated by set theory. Then we will show that triangle inequality holds true for the corresponding distance measure defined on general set space, which will be discussed by Lemma B.2-B.6, At last, through Theorem B.2, we would show that triangle inequality holds true for the graph distance measure defined in Definition 4.3.

Obviously, we have $d_f(G_1,G_2)=$ $d_f(G_1',G_2')$ if $G_1 \cong G_1'$ and $G_2 \cong G_2'$. Then we can assume that *for any graphs under consideration, the vertex sets of these graphs are pairwise disjoint.* In other words, for any graphs sharing common vertexes, we can find corresponding isomorphic graphs without common vertexes, such that the quantification relation between original graphs is conserved in these isomorphic copies.

Let U be an universe vertex set, then we denote by $\tilde{G}(V)$ the set consisting of all graphs with vertex set $V \subseteq U$, *i.e.* $\tilde{G}(V)=\{G|V(G)=V\}$. And we use $G^*(V)$ to denote the set of subgraphs of $\tilde{G}(V)$, *i.e.* $G^*(V)=\{G|V(G) \subseteq V\}$. Obviously, given any vertex set V , we can get a graph class $G^*(V)$

Lemma B.1. For any two graphs G_1 and G_2 , *w.l.o.g* let $V(G_1) \cap V(G_2)=\varnothing$, let $\boldsymbol{H} =\{G_1', G_2'\} \subseteq G^*(V)$, then there exist $V \subseteq U$ with cardinality as $|V(G_1)|+|V(G_2)|\text{-}|V(G_{12})|$, such that (1) $G_i \cong G_i'(i=1,2)$ and (2) $G_{12} \cong G_1' \cap G_2'$, where G_{12} is the maximum common edge induced graph between G_1 and G_2, and G_{12} is not necessarily to be non-empty graph.

Proof: It's clear that if $G_{12}=\varnothing$, the theorem holds true. If $G_{12} \neq \varnothing$, Let $g_1 \subseteq G_1$, $g_2 \subseteq G_2$ and $g_1 \cong g_2 \cong G_{12}$, then there is a one-to-one mapping ϕ: $V(g_1) \rightarrow V(g_2)$, such that the adjacent relations are conserved between g_1 and g_2. Let U be a vertex set disjoint to $V(G_1) \cup V(G_2)$, such that $|U| >> |V(G_1)|+|V(G_2)|$. Then we can construct a mapping γ from $V(G_1) \cup V(G_2)$ to U as follows.

Let $V \subseteq U$ and $|V|=|V(G_1)|+|V(G_2)|\text{-}|V(G_{12})|$. We construct a *partition* $\tilde{V}=\{V_1,V_2,V_3\}$, such that $|V_1|=|V(G_1)\text{-}V(g_1)|$, $|V_2|=|V(G_2)\text{-}V(g_2)|$ and $|V_3|=|V(g_1)|$.

Thus we can easily construct two *bijective* mapping γ_1: $V(G_1)\text{-}V(g_1) \rightarrow V_1$ and γ_2: $V(G_2)\text{-}V(g_2) \rightarrow V_2$. We also can construct a mapping γ_3: $V(g_1)\cup V(g_2)\rightarrow V_3$, such that $\forall v\in V(g_1)$, $\gamma_3(v)= \gamma_3(\phi(v))$. Then γ can be constructed as follows:

$$\gamma(v)=\begin{cases}\gamma_1(v) & v\in V(G_1)\text{-}V(g_1)\\ \gamma_2(v) & v\in V(G_2)\text{-}V(g_2)\\ \gamma_3(v) & v\in V(g_1)\cup V(g_2)\end{cases}$$

Obviously, we can find a graph $g\in \tilde{G}(V_3)$ *s.t.* $g\cong g_1\cong g_2\cong G_{12}$. Then we need to construct two graphs G_1', G_2' in terms of vertex set V, *i.e.* G_1', G_2'$\in G^*(V)$. We only need to show how to construct G_1' with vertex set $\gamma(V(G_1))$ in the following way. First let G_1'$=(\gamma(V(G_1)), \varnothing)\cup g$, then $\forall(u, v)\in E(G_1)$, let $E(G_1$'$)= E(G_1$'$)\cup(\gamma (u), \gamma (v))$. It's not difficult to show that G_1' constructed in this way is isomorphic to G_1. Similarly, we can construct G_2' such that G_2'$\in G^*(V)$ and G_2'$\cong G_2$.

Then we need to show that $G_{12}\cong G_1$'$\cap G_2$' . From the construction process of G_1', G_2' , we have $g\subseteq G_1$' and $g \subseteq G_2$', thus $g\subseteq G_1$'$\cap G_2$' Assume G_{12} is not isomorphic to G_1'$\cap G_2$', then it follows that G_1'$\cap G_2$'- $g\neq\varnothing$. Obviously, G_1'$\cap G_2$' is a common graph larger than g , due to G_1'$\cong G_1$, G_2'$\cong G_2$ and $g\cong G_{12}$, then we get the conclusion that G_1 and G_2 have a larger common graph than G_{12}, which contradicts to the condition that G_{12} is a maximum common graph of G_1 *and* G_2.

An immediate consequence of Lemma B.1 is the following Theorem B.1.

Theorem B.1. For any three graphs G_1, G_2 and G_3 , let $\mathbf{H} =\{ G_1$', G_2' G_3' $\}\subseteq G^*(V)$, then there exist $V\subseteq U$ with cardinality as $|V(G_1)|+|V(G_2)|+|V(G_3)|-|V(G_{12})|-|V(G_{12})|-|V(G_{12})|+|V(G_{123})|$, such that (1) $G_i\cong G_i$' ($i=1,2,3$) and (2) $G_{ij}\cong G_i$'$\cap G_j$' for any i and j ($i,j=1,2,3$), where G_{ij} is the maximum common edge induced graph between G_i and G_j, and G_{ij} is not necessarily to be non-empty graph*;* G_{123} is the maximum common edge induced graph between G_1, G_2 and G_3, G_{123} is not necessarily to be non-empty graph.

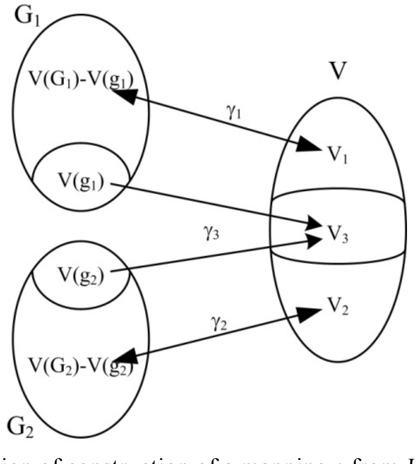


Figure B.1: Illustration of construction of a mapping γ from $V(G_1)\cup V(G_2)$ to U

Example B.1 We use this example to illustrate the transformation of graph space discussed in Lemma B.1 and Theorem B.1. Given arbitrary two graphs G_1 and G_2, as shown in Figure B.2, we use *f*(*fv*,*fe*)[5] to denote the isomorphism corresponding to the maximum common edge induced graph between G_1 and G_2, where *fv*={(1,1),(2,2),(3,3),(4,4)} and *fe*={((12),(12)), ((13),(13)), ((23),(23)), ((24),(24)), ((34),(34))}. If G_1 and G_2 come

[5] The isomorphism corresponding to the maximal common edge induced graph between G1 and G2 can be determined by a pair of mapping *f*(*fv*, *fe*), where *fv* is the vertex mapping from $V(G_1)$ into $V(G_2)$, *fe* is the edge mapping from $V(G_1)$ into V(G2). We use ordered pair (v_1,v_2), where $v_1\in V(G_1)$ and $v_2\in V(G_2)$, to represent *fv*, use ordered pair (e_1,e_2), where $e_1\in E(G_1)$ and $e_2\in E(G_2)$, to represent *fe*.

from the same vertex set, then by the assumption discussed as before, we can found two graphs G_1' and G_2' sharing no common vertexes, and isomorphic to G_1 and G_2 respectively. Furthermore, we can construct two graphs G_1'' and G_2'' with vertex set U_3, such that $G_1'\cong G_1''$ and $G_2'\cong G_2''$.Obviously, the join of G_1'' and G_2'' is just isomorphic to the maximum common edge induced subgraph of G_1 and G_2

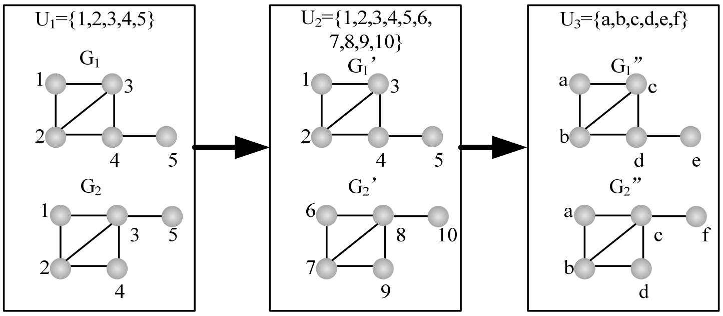


Figure B.2: Illustration of space transformation for any given arbitrary graph space

Thus, we have constructed an isomorphic graph space for arbitrary three graphs G_1, G_2 and G_3. The significance of the transformation described in Lemma B.1 and Theorem B.1 lies in the fact that in the original graph space, the maximum common subgraph is worked out by looking for a maximum graph isomorphism of graphs, whereas in the transformed graph space, the maximum common subgraph or maximum graph isomorphism is equivalent to the join of corresponding graphs that isomorphic to the original graphs. Consequently, we can discuss graph isomorphism related problems in the context of traditional set theory.

The following lemmas (Lemma B.2 to B.6) will discuss triangle inequality defined on set space. We begin this section with some basic notations. Let X be a set consisting n elements, *i.e.* $X=\{x_i|i=1,\ldots,n\}$, let $X^k=\{B|B\subseteq X$ and $|B|=k\}$. Let $\Omega(X)$ be the set of all subset of X.

Lemma B.2. Let X be any set, let $d(A,B)=1-|A\cap B|/Max(|A|,|B|)$ be a distance measure defined on $\Omega(X)$, then for any three set $A, B, C\in\Omega(X)$, triangle inequality holds true for $(\Omega(X),d)$.

Lemma B.3. For two sets $A,B\subseteq X$, if $A\subseteq B$, then

(1) $A^k\subseteq B^k$, where $k\leq|A|$.

(2) $A\in\Omega(B)$

Lemma B.4. For two sets $A,B\subseteq X$, the following statements hold true:

(1) *(a)* $A^k\cap B^k=(A\cap B)^k$; *(b)* $A^k\cup B^k\subseteq(A\cup B)^k$, where $k\leq|X|$

(2) *(a)* $A^I\cap B^I=(A\cap B)^I$; *(b)* $A^I\cup B^I\subseteq(A\cup B)^I$, where $I\subseteq J$ and $J=\{1,2,3...min(|A|,|B|)\}$,

(3) $A^i\cap B^j=\varnothing$, for any pair (i, j), $i\neq j$

The proof of Lemma B.2 is similar to the proof in Appendix A, it is omitted in this section. The details of the proof of Lemma B.3 and Lemma B.4 are not very difficult; we leave them to the reader.

Lemma B.5. Let X be any set, for any three sets $A, B, C\subseteq X$, triangle inequality holds true for $(\Omega(X),d_i)$, where $d_i(A,B)= 1-|A^i\cap B^i|/Max(|A^i|, |B^i|)$ is a distance measure defined on $\Omega(X)$.

Proof: For any three set $A, B, C\subseteq X$, we have $A^i, B^i, C^i\subseteq X^i$,then $A^i, B^i, C^i\in X^i$ .We can define a distance measure on

$\Omega(X^i)$ as $d(A^i,B^i)= 1\text{-}|A^i\cap B^i|/Max(|A^i|,|B^i|)$. According to Lemma B.2, triangle inequality holds true for $(\Omega(X^i)\ ,d)$. Thus, we also have that triangle inequality holds true for $(\Omega(X),\ d_i)$, where $d_i(A,B)= 1\text{-}|A^i\cap B^i|/Max(|A^i|,\ |B^i|)$.

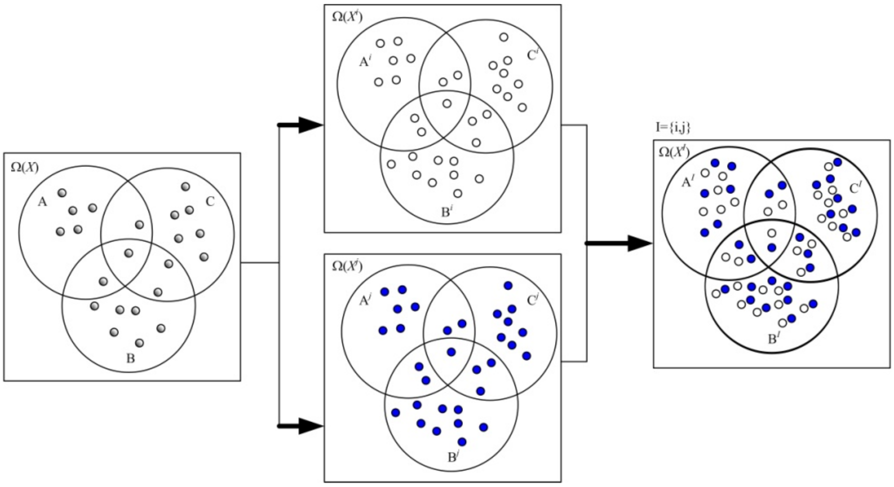


Figure B.3: Illustration of three space $(\Omega(X),d)$,$(\Omega(X^i),d)$, $(\Omega(X^j),d)$ and $(\Omega(X^I),d)$

Lemma B.6. Let X be any set, let $J=\{1,2,3...|X|\}$, for any three sets $A,\ B,\ C\subseteq X$, then triangle inequality holds true for distance measure: $d_I(A,B)= 1\text{-}|A^I\cap B^I|/Max(|A^I|,\ |B^I|)$, where $A^I=\cup_{(i\in I)}A_i$ and $I\subseteq J$.

Proof: For any three sets $A,\ B,\ C\subseteq X$, it follows that for each $i\in I$, $A^i,\ B^i,\ C^i\subseteq X^i$, thus we have $A^I,\ B^I,\ C^I\subseteq X^I$, then $A^I,\ B^I,\ C^I\in\Omega(X^I)$. Thus we can define a distance measure on $\Omega(X^I)$ as $d(A^I,B^I)= 1\text{-}|A^I\cap B^I|/Max(|A^I|,|B^I|)$. According to Lemma B.2, triangle inequality holds true for $(\Omega(X^I)\ ,d)$. Thus, we also have that triangle inequality holds true for $(\Omega(X)\ ,d_I)$, where $d_I(A,B)= 1\text{-}|A^I\cap B^I|/Max(|A^I|,\ |B^I|)$.

Example B.2 As shown in Figure B.3, let X be any set, from Lemma B.1, it follows that triangle inequality holds true for $(\Omega(X),d)$, $(\Omega(X^i),d)$, $(\Omega(X^j),d)$ and $(\Omega(X^I),d)$ where d follows the form defined in Lemma B.1

Theorem B.2. For any three graphs G_1, G_2 and G_3 , let $J=\{1,2,3...|X|\}$, $I\subseteq J$, then triangle inequality holds true for distance measure: $d_I(G_1,G_2)=1\text{-}|\Gamma_I\ (G_{12})|/Max(|\Gamma_I\ (G_1)|,|\Gamma_I(G_2)|)$, where $\Gamma^I=\cup_{(i\in I)}\ \Gamma_i$.

Proof: From theorem B.1, for graphs G_1, G_2 and G_3, we can construct G_1', G_2' and G_3' with certain vertex set V, such tha*t* (1) $G_i\cong G_i'(i=1,2,3)$ and (2) $G_{ij}\cong G_i'\cap G_j'$ for any i and j $(i,j=1,2,3)$.

Clearly, it follows that $d_I(G_i,G_j)=1\text{-}|\Gamma_I\ (G_{ij})|/Max(|\Gamma_I\ (G_i)|,|\Gamma_I(G_j)|)=1\text{-}|\Gamma_I\ (G_i'\cap G_j')|/Max(|\Gamma_I\ (G_i')|,|\Gamma_I(G_j')|)= d_I(G_i',G_j')=d_I(E_i',E_j')$. From Lemma B.6, we have that triangle inequality holds true for $(\Omega(V\times V),\ d_I)$. Hence we can conclude that triangle inequality holds true for d_I defined on graph space.