Graph Kernels for Chemoinformatics

Ramzi CHAIEB

Research Laboratory:

LITIS- Laboratoire d'Informatique, du Traitement de l'Information et des Systèmes

ramzi.chaieb@univ-rouen.fr

03/05/2019

Plan

1. Introduction and objectives

2. New molecular graph representation

3. New molecular descriptors (calculated from the new dataset)

4. Comparison of molecular graphs

5. New graph kernels

6. Conclusion

Introduction and objectives

Introduction and objectives

- Chemoinformatics is the application of computer science to problems related to chemistry

- Problems treated by chemoinformatics:

  ✓ Representation of molecules and management of molecular databases

  ✓ Prediction of physical or biological properties of molecules

  ✓ Drug design

  ✓ Resolution of molecular structures

  ✓ Prediction of chemical reactions

Introduction and objectives

- Chemoinformatics uses methods derived from computer science: graph theory and machine learning

→ **Classify** or **predict** the basic properties of molecules

- Graph kernels provide an interesting approach by combining **automatic learning methods** and **the natural representation of molecules by graphs**

- Several methods based on graphical kernels have been proposed to solve problems in chemoinformatics

Introduction and objectives

- Contribute to the problem of molecular representation and the measure of similarity between molecular graphs

➔ New molecular representation encoding local electronic information

➔ New similarity measure as a kernel for comparing two molecules encoded in the new proposed representation

➔ Optimization of graph kernels

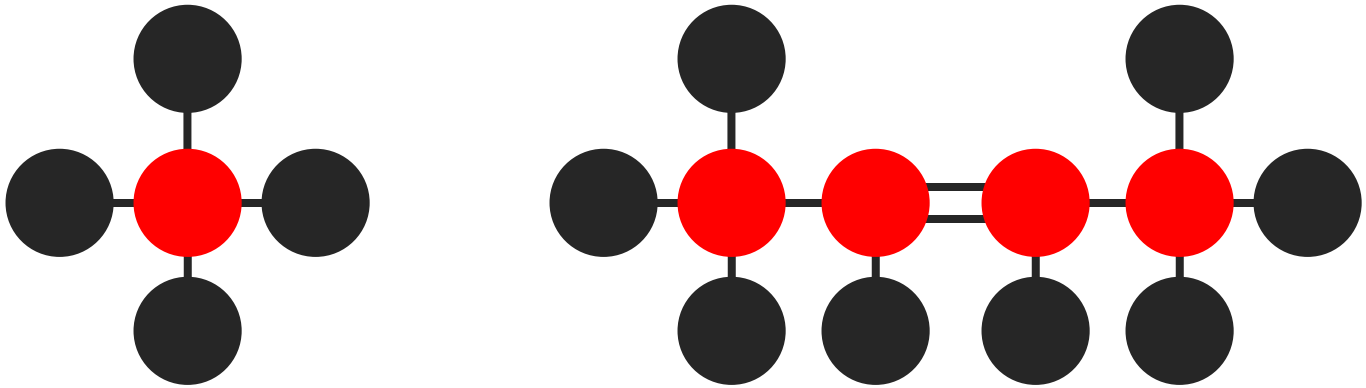
2

New molecular graph representation

Molecular graph

- The molecular graph is a simple, labeled, unoriented graph representing the structure of a molecule

- The set of vertices encodes the atoms
- The set of edges represents the covalent bonds between the atoms



CH_4 C_4H_8

Figure 1: Examples of molecules

Molecular graph

- Each vertex is labeled by:
  - ✓ The chemical element of the corresponding atom
  - ✓ A vector of the intrinsic characteristics of each atom
- Each edge is labeled by:
  - ✓ The type of covalent bond (single, double, triple or aromatic)

The molecular graph encodes the adjacency relations between atoms but not their geometrical positions (Euclidean distance between atoms).

Databases of molecular graphs

- The characteristics of each database of molecular graphs are:

✓ The number of molecules

✓ The number of average vertices of the molecular graphs representing the molecules

✓ Their average degree as well as the extremes (Minimum size and Maximum size) of the sizes of the molecular graphs

Databases of molecular graphs

- **Types of problem (Classification or Regression)**

✓ **Classification problem**

→ Predict the class of a molecule

✓ **Regression problem**

→ Predict a property, usually physical, that can take real values

(example: Prediction of the boiling temperature of molecules, ...)

New molecular descriptors calculated from the new dataset

New molecular descriptors calculated from the new dataset

- Connectivity Matrices
- Global Quantum Molecular Descriptors
- Atomic Quantum Molecular Descriptors
- “Composite” Quantum Molecular Descriptors
- Summary of extracted descriptors

New molecular descriptors calculated from the new dataset

Connectivity Matrices

C1	C2	(Bonded)
C1	C3	(Bonded)
C3	N4	(Bonded)
C2	N5	(Bonded)
C1	C6	(Bonded)
C6	H7	(Bonded)
C6	C8	(Bonded)
C8	C9	(Bonded)
C8	C10	(Bonded)
C9	C11	(Bonded)
C9	H12	(Bonded)
C10	C13	(Bonded)
C2	H14	(Bonded)
C10	H14	(Bonded)
C11	C15	(Bonded)
C13	C15	(Bonded)
C11	H16	(Bonded)
C13	H17	(Bonded)
C15	H18	(Bonded)

→

	C1	C2	C3	N4	N5	C6	H7	C8	C9	C10	C11	H12	C13	H14	C15	H16	H17	H18
C1	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
C2	1	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
C3	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N4	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N5	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
C6	1	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0
H7	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
C8	0	0	0	0	0	1	0	1	1	1	0	0	0	0	0	0	0	0
C9	0	0	0	0	0	0	0	1	1	0	1	1	0	0	0	0	0	0
C10	0	0	0	0	0	0	0	1	0	1	0	0	1	1	0	0	0	0
C11	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	1	0	0
H12	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0
C13	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	1	0
H14	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
C15	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	1
H16	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0
H17	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0
H18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1

Table 1: Example of a connectivity matrix (The first molecule: 1_MV)

New molecular descriptors calculated from the new dataset

- **Global Quantum Molecular Descriptors**

✓ These descriptors correspond to one value that describes the whole molecule

✓ Some of them are not really useful, but they may be needed in order to calculate other descriptors

16 descriptors:

μ^{+}_{FMO}
μ^{-}_{FMO}
μ_{FMO}
η_{FMO}
S_{FMO}
γ_{FMO}

ω_{FMO}
ω^{+}_{FMO}
ω^{-}_{FMO}
ΔE^{nucleo}_{FMO}
$\Delta E^{electro}_{FMO}$

$S^{(2)}_{FMO}$
$\Delta\omega^{\pm}{}_{FMO}$
ΔN^{Max}_{FMO}
P_{FMO}
IRI_{FMO}

New molecular descriptors calculated from the new dataset

- **Atomic Quantum Molecular Descriptors**

✓ These descriptors will help to create the atomic matrices

✓ This extraction should be done on all the atoms

29 descriptors:

q_{C1}
K_{C1}
K_Scaled_{C1}
Mu_Intra_{C1}
ESP_{C1}
$ESPe_{C1}$
$ESPn_{C1}$
E_IQA_{C1}
V_IQA_{C1}
VC_IQA_{C1}
VX_IQA_{C1}

$E\_IQA\_Intra_{C1}$
$V\_IQA\_Intra_{C1}$
$VC\_IQA\_Intra_{C1}$
$VX\_IQA\_Intra_{C1}$
LI_{C1}
$DI(A,A')_{C1}$
$ESP\_Max\_IDS_{C1}$
$ESP\_Min\_IDS_{C1}$
$ESP\_Avg\_IDS_{C1}$
Vol_{C1}

$T\_Q(A)_{C1}$
$R+2_{C1}$
EhF_{C1}
KE_Weiz_{C1}
Ke_TF_{C1}
$f^{+}{}_{FMO,C1}$
$f^{-}{}_{FMO,C1}$
$f^{(2)}{}_{FMO,C1}$

New molecular descriptors calculated from the new dataset

- **"Composite" Quantum Molecular Descriptors**

✓ They are atomic descriptors that are usually a combination of 2 descriptors, one global with an atomic one

21 descriptors:

$\mu^{+}_{FMO} f^{+}_{FMO,C1}$
$\mu^{-}_{FMO} f^{-}_{FMO,C1}$
$\omega_{FMO} f^{+}_{FMO,C1}$
$\omega_{FMO} f^{-}_{FMO,C1}$
$\omega^{+}_{FMO} f^{+}_{FMO,C1}$
$\omega^{-}_{FMO} f^{-}_{FMO,C1}$
$\omega_{FMO} f^{(2)}_{FMO,C1}$
$s^{+}_{FMO,C1}$
$s^{-}_{FMO,C1}$
$S_{FMO} f^{(2)}_{FMO,C1}$

$s^{+}_{K\,FMO,C1}$
$s^{-}_{K\,FMO,C1}$
$S^{(2)}_{FMO} f^{(2)}_{FMO,C1}$
$s^{(2)}_{FMO,C1}$
$\Delta\rho^{+}_{\ FMO,C1}$
$\Delta\rho^{-}_{\ FMO,C1}$
$\mu_{FMO,C1}$
$\eta_{FMO,C1}$
$P_{FMO} f^{+}_{FMO,C1}$
$P_{FMO} f^{-}_{FMO,C1}$
$P_{FMO} f^{(2)}_{FMO,C1}$

New molecular descriptors calculated from the new dataset

- **Summary of extracted descriptors**

66 descriptors:

μ^{+}_{FMO} μ^{-}_{FMO} μ_{FMO} η_{FMO} S_{FMO} γ_{FMO}

ω_{FMO} ω^{+}_{FMO} ω^{-}_{FMO} ΔE^{nucleo}_{FMO} $\Delta E^{electro}_{FMO}$

$S^{(2)}_{FMO}$ $\Delta\omega^{\pm}_{FMO}$ ΔN^{Max}_{FMO} P_{FMO} IRI_{FMO}

q_{C1} K_{C1} K_Scaled_{C1} Mu_Intra_{C1} ESP_{C1} $ESPe_{C1}$ $ESPn_{C1}$ E_IQA_{C1} V_IQA_{C1} VC_IQA_{C1} VX_IQA_{C1}

$E\_IQA\_Intra_{C1}$ $V\_IQA\_Intra_{C1}$ $VC\_IQA\_Intra_{C1}$ $VX\_IQA\_Intra_{C1}$ LI_{C1} $DI(A,A')_{C1}$ $ESP\_Max\_IDS_{C1}$ $ESP\_Min\_IDS_{C1}$ $ESP\_Avg\_IDS_{C1}$ Vol_{C1}

$T\_Q(A)_{C1}$ $R+2_{C1}$ EhF_{C1} KE_Weiz_{C1} Ke_TF_{C1} $f^{+}_{FMO,C1}$ $f^{-}_{FMO,C1}$ $f^{(2)}_{FMO,C1}$

$\mu^{+}_{FMO}f^{+}_{FMO,C1}$ $\mu^{-}_{FMO}f^{-}_{FMO,C1}$ $\omega_{FMO}f^{+}_{FMO,C1}$ $\omega_{FMO}f^{-}_{FMO,C1}$ $\omega^{+}_{FMO}f^{+}_{FMO,C1}$ $\omega^{-}_{FMO}f^{-}_{FMO,C1}$ $\omega_{FMO}f^{(2)}_{FMO,C1}$ $s^{+}_{FMO,C1}$ $s^{-}_{FMO,C1}$ $S_{FMO}f^{(2)}_{FMO,C1}$

$s^{+}_{K\,FMO,C1}$ $s^{-}_{K\,FMO,C1}$ $S^{(2)}_{FMO}f^{(2)}_{FMO,C1}$ $s^{(2)}_{FMO,C1}$ $\Delta\rho^{+}_{FMO,C1}$ $\Delta\rho^{-}_{FMO,C1}$ $\mu_{FMO,C1}$ $\eta_{FMO,C1}$ $P_{FMO}f^{+}_{FMO,C1}$ $P_{FMO}f^{-}_{FMO,C1}$ $P_{FMO}f^{(2)}_{FMO,C1}$

Comparison of molecular graphs

Comparison methods based on vector representations

- Encode each molecular graph by a vector of fixed size

✓ **Local approaches**

Each element of the vector encodes:

- The number of occurrences of a tag of a node
- The number of occurrences of each edge

A histogram encodes the distribution of node and edge labels

✓ **Global approaches**

Encode the graph in a global way (global information such as the number of nodes or edges of the graph.)

Comparison methods based on vector representations

\+ Use most machine learning algorithms

\- Induce a loss of information by coding the graph by a fixed size vector

→ To avoid or limit this loss of structural information, other methods deduce the similarity of the molecular graphs based directly on the graph

Comparison methods based on vector representations

- Graphs provide a generic data structure widely used in chemo and bioinformatics to represent complex structures such as chemical compounds or complex interactions between proteins

- The high flexibility of this data structure does not allow to readily combine it with usual machine learning algorithms based on a vectorial representation of input data

- Graph matching algorithms are NP-hard to compute

Comparison methods based on vector representations

→ Graph embedding methods aims to tackle this limitation by embedding graphs into explicit vectorial representations which allow the use of any machine learning algorithm defined on vectorial representations

- However, encoding graphs as explicit vectors having a limited size induces a loss of information which may reduce prediction accuracy

Vector representation / graph-based representation

	Vector representation	Graph-based representation
Advantages	• Use a wide variety of mathematical tools defined in a vector space. • Quick	• Exploit a maximum of structural information encoded in molecular graphs.
Disadvantages	• Encode a limited set of information. • Loss of structural information	• Can not directly apply the majority of machine learning algorithms. • Exponential complexity

Table 2: Comparison between a vector representation and a graph-based representation

New graph kernels

New graph kernels

→ Graph kernels can be understood as graph similarity measures corresponding to scalar products between graph's projections in an implicit (and possibly unknown) Hilbert space

+ Graph kernels can be used in any kernel method thanks to the underlying implicit embedding

+ Relaxing the constraints associated to explicit vectorial representations of graphs limits the loss of structural information

→ Graph kernels allow to encode more structural information than classic graph embedding methods while being able to connect to powerful machine learning methods

New graph kernels

- Graph kernels have been defined on molecular graphs and applied to chemoinformatics to build prediction models using the structural information encoded within molecular compounds

- However, intrinsic properties of atoms and theirs interactions induce some electronic properties which are not explicitly encoded within classic molecular graphs representations

→ Include intrinsic properties of atoms and theirs interactions into a new augmented kernel and apply it on some chemoinformatics datasets

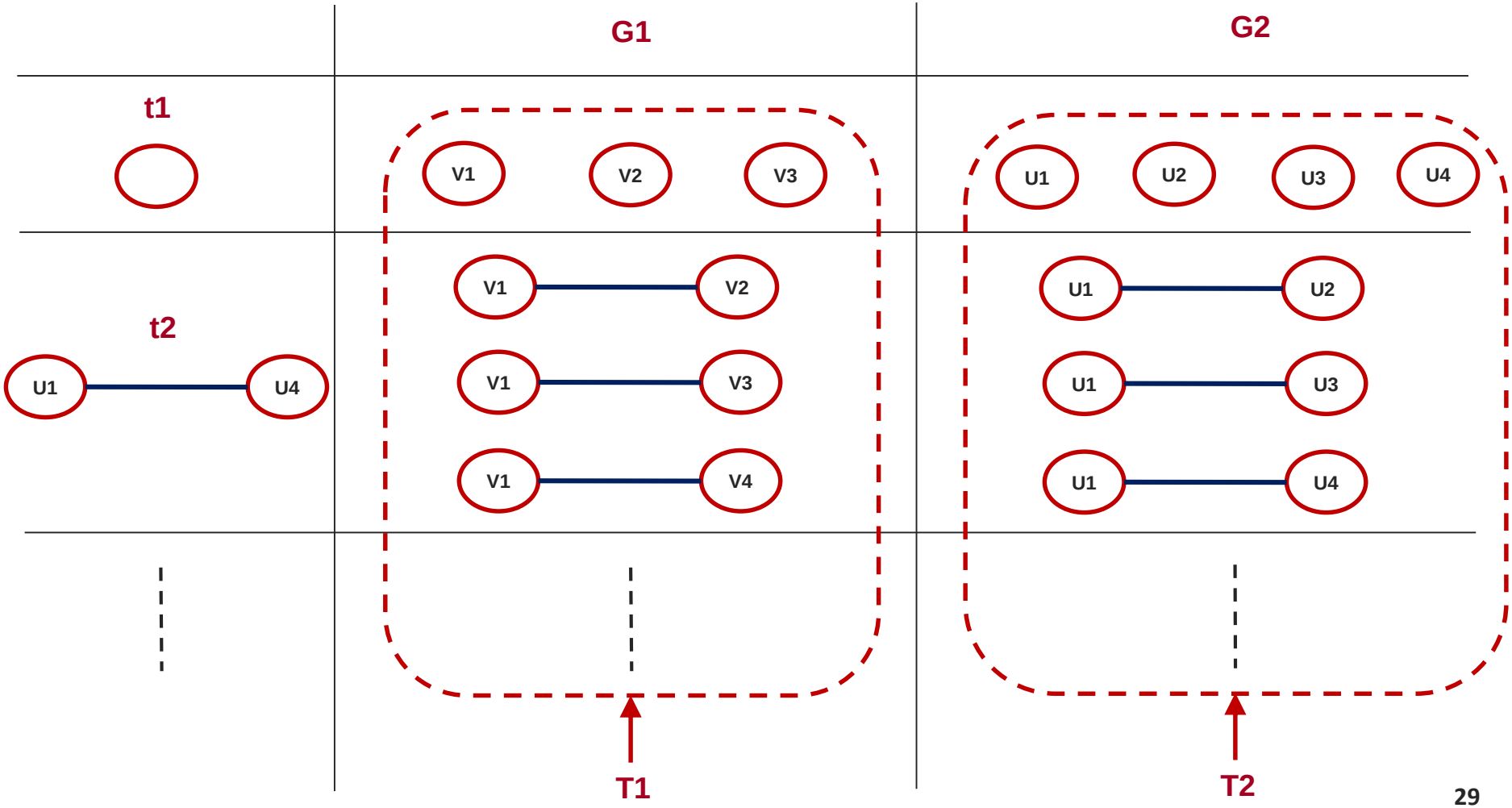
New graph kernels

First proposal:

Figure 2: Example of two graphs

New graph kernels

- **Step 1**
  - ✓ Search all possible patterns in G1 and G2



New graph kernels

- **Step 2**
  - ✓ Calculate the kernel K(G1, G2) between G1 and G2

$$K(G1,G2)=\sum_{t\in T1}\sum_{t'\in T2}K_{\varphi}(t,t')$$

kernel between G1 and G2

$$K_{\varphi}(t,t')=\begin{cases}0\\ K_{\sigma}(t,t')\end{cases}$$

If pattern(t) ≠ pattern(t')

If pattern(t) = pattern(t') **without taking into account the labels of the nodes and the edges**

Kernel between 2 substructures

New graph kernels

$$K_\sigma(t,t') = \prod_{i=1\ldots|t|} k_v(t(i),t'(i)) * \prod_{i=1\ldots|t|} k_e((t(i),t(i+1),(t'(i),t'(i+1))$$

Similarity between 2 substructures that have the same form

Kernel between the nodes

Kernel between the edges

New graph kernels

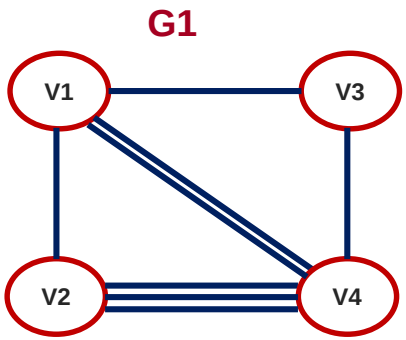
Second proposal :

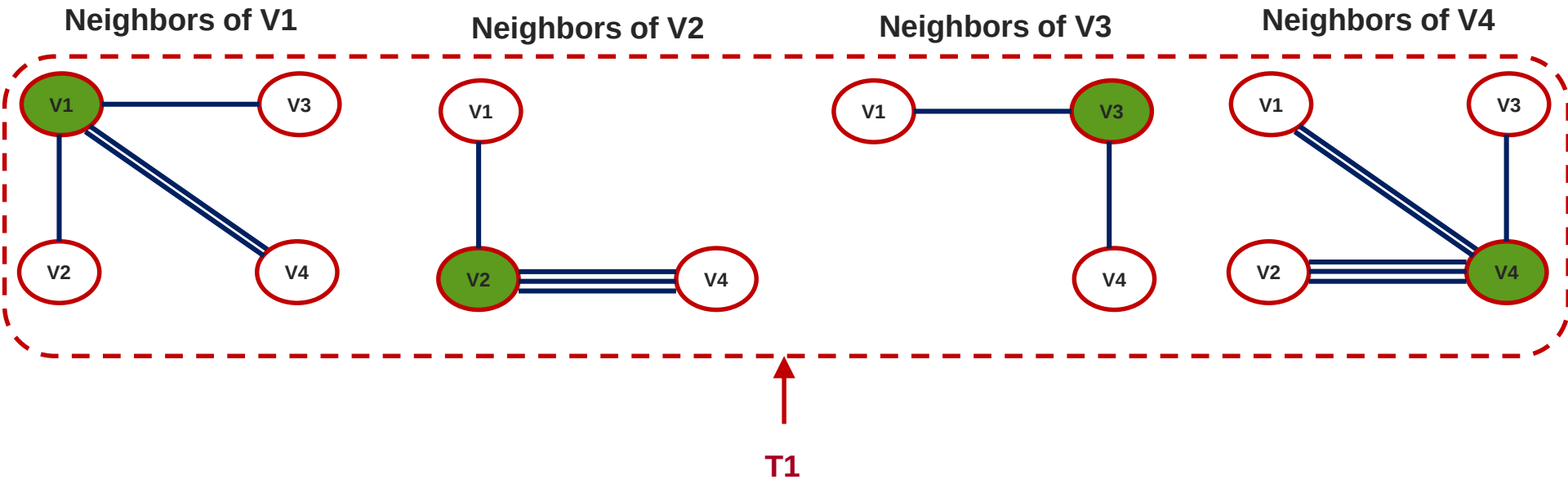
Figure 3: Example of two graphs

New graph kernels

- **Step 1**
  - ✓ Find all the neighbors of depth p for each node in G1 and G2

For p = 1

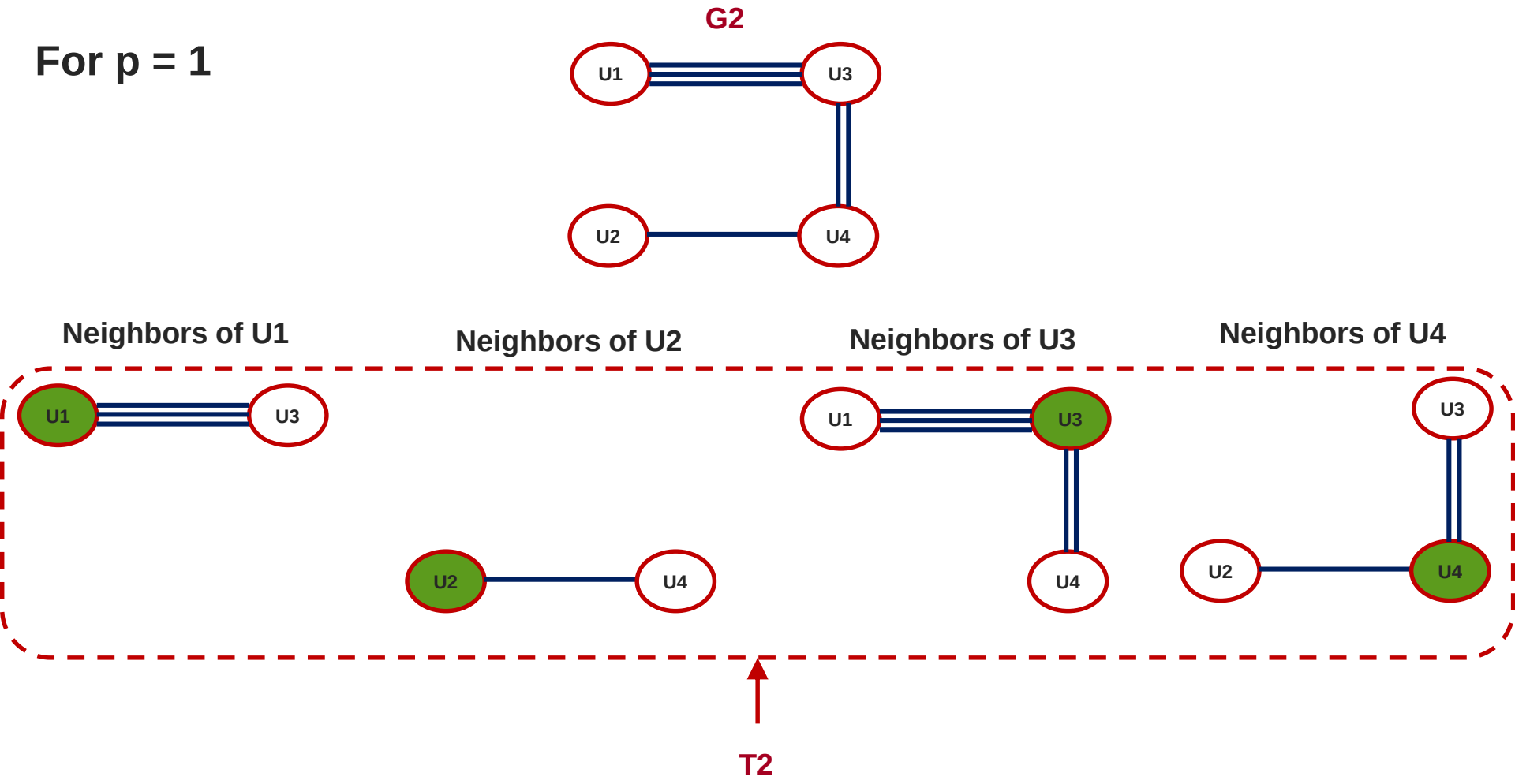




New graph kernels

- **Step 1**
  - ✓ Find all the neighbors of depth p for each node in G1 and G2

For p = 1



New graph kernels

- **Step 2**
  - ✓ Calculate the kernel K(G1, G2) between G1 and G2

$$K(G1,G2) = \sum_{t \in T1} \sum_{t' \in T2} K_{\varphi}(t,t')$$

Kernel between G1 and G2

$$K_{\varphi}(t,t') = w_{np} \prod_{i=1...\max(|t|)} k_{np}(t(i),t'(i)) * w_{nv} \prod_{i=1...\max(|t|)} k_{nv}(t(i),t'(i)) * w_{e} \prod_{i=1...\max(|t|)} k_{e}((t(i),t(i+1)),(t'(i),t'(i+1)))$$

Similarity between 2 substructures

Kernel between neighboring nodes

Kernel between the main nodes

Kernel between the edges

$$w_{np} + w_{e} + w_{nv} = 1$$

Conclusion

Conclusion

- New molecular representation encoding local electronic information

  ✓ Extract the maximum of information from a molecule

  ✓ Exploit the advantages of vector and structural representations

- New kernels to measure the similarity between two molecular graphs

- Apply the new augmented kernel on some chemoinformatics datasets

Thank you