

# Developing and evaluating a Cytoscape app for graph-based clustering

Philipp Spohr

September 9, 2017

**Abstract**

**Contents**

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>The Yoshiko-App for Cytoscape</b>	<b>3</b>
2.1	Technical Details . . . . .	3
2.2	Algorithm . . . . .	3
2.2.1	Data Modeling . . . . .	3

# 1 Introduction

## 2 The Yoshiko-App for Cytoscape

### 2.1 Technical Details

### 2.2 Algorithm

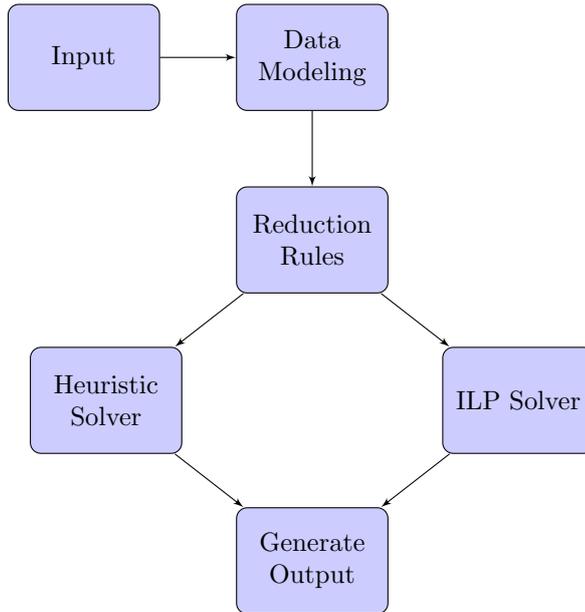


Figure 1: Overview of the Yoshiko Algorithm

ROUGH IDEA: COMPLETE GRAPH, CHOOSE EDGES SO THAT sum of  $C(E)$  is MAX while satisfying Triangle inequalities  $\wr$  Fully Disjunct Clique-Graph

#### 2.2.1 Data Modeling

**Theory** The Yoshiko algorithm models the data as a complete graph  $G = (V, E)$  with an associated edge-cost function  $C : E \rightarrow \mathbb{R} \cup \{-\infty\} \cup \{\infty\}$ . As many input instances do not describe a full graph, missing edges and costs need to be modeled. This is achieved by using default values for insertion or deletion. A default insertion cost  $C_I \in [-\infty, 0]$  is used as  $C(e)$  whenever the input instance does not contain an edge  $e$ . A default deletion cost  $C_D \in [0, \infty]$  is used as  $C(e)$  whenever the input instance does contain an edge  $e$  that has no cost associated yet.

**Implementation** The Yoshiko Wrapper provides a clean and simple interface to generate the model.

**Mapping edge costs** The user has the possibility to use a numeric Cytoscape column of the node table as a source for the edge-cost function  $C$ .

**Insertion and deletion cost** The default values  $C_I$  and  $C_D$  can be set by the user with the default values being  $C_I = -1$  and  $C_D = 1$ . It should be noted, that the insertion cost value is not normalized or in any way adjusted when a mapping is used. This means that the user needs to adjust this value wisely to fit the data. As an example the user might have mapped the edge costs to a column containing values in the range of  $10^6 - 10^7$ . The default insertion cost of  $-1$  will be irrelevant in comparison and the algorithm will most likely insert all missing edges and generate one big cluster as a solution. Overall the ratio  $R = \frac{|C_I|}{C_D}$  should give you a rough idea how the algorithm will operate.  $R > 1$  means, that the algorithm is more likely to delete edges to generate cliques, a value of  $R < 1$  means insertions are more likely.

**Mapping permanent or forbidden edges** The Yoshiko app has additional convenience functions: The user can map edges to a boolean Cytoscape column to mark them as either **forbidden** (meaning that those edges will never be part of the solution) or **permanent** (meaning that those edges will always be part of the solution). Marking an edge  $e$  as forbidden is equivalent to  $C(e) = -\infty$ , marking an edge  $e$  as permanent is equivalent to  $C(e) = \infty$ . This way the user is able to apply expert knowledge about the input instance to increase the quality of the solution.