

www.qconferences.com  
www.qconbeijing.com



伦敦 | 北京 | 东京 | 纽约 | 圣保罗 | 上海 | 旧金山

London · Beijing · Tokyo · New York · Sao Paulo · Shanghai · San Francisco

# QCon全球软件开发大会

International Software Development Conference



Applying graph mining and machine learning to  
detect fraudulent networks

图数据挖掘技术在PayPal风险管理中的研究与应用

饶卫雄

[rweixiong@gmail.com](mailto:rweixiong@gmail.com)

2013.4.27



# Motivation背景

- Online Payment System suffers from fraudulent transactions, which result in the loss of customers and service providers.
- 在线支付系统一旦产生欺诈交易，导致客户与服务供应商的经济损失
- In common scenarios, those transactions occur with patterns:
- 欺诈交易具有如下模式:
  - One account is hacked into by criminals, then multiple fraudulent transactions are submitted to transfer money from that account to somewhere else.
  - 犯罪分子盗窃在线帐户，通过欺诈交易进行资金转帐。
  - Illegal accounts are created and linked to a stolen (or fake) credit card, and then fraudulent transactions are submitted to withdraw money from those accounts.
  - 关联非法创建的帐户与盗窃或者虚假的信用卡，通过欺诈交易从该帐户中进行转移资金的。
  - Customers' computers are affected by Special designed virus, and then multiple fraudulent transactions are submitted to transfer their money into one account.
  - 大量的客户计算机被病毒感染之后，欺诈交易会将其客户资金转移到特定帐户。
- Every pattern can be seen as a set of accounts and transactions, where those transactions or accounts have strong relations with each other. (e.g. Similar transaction time, Same IP Address, and so forth)
- 每个模式均可以根据在线帐户与交易的集合纪录进行抽象特征提取。



同濟大學  
TONGJI UNIVERSITY

PayPal™

# Objective 目标

- We build a transaction graph (or network) based on the relationships of accounts and transactions. So all the historical data can be linked to generate the universal transaction graph.
- 通过帐户和交易的关联关系，构建交易图(或者资金转移网络)，所有的历史数据均关联至该交易图。
- By adapting graph mining techniques, we can identify fraudulent networks, which exist as sub-graphs of the universal transaction graph.
- 使用图挖掘技术，识别交易图中的欺诈网络子图。
- By adapting machine learning algorithms, we can cluster those fraudulent networks and generate fraudulent transaction patterns.
- 使用机器学习算法，进一步鉴别新式的欺诈交易模式。
- Finally, with all the models and information above, we can recognize whether a unknown transaction is fraudulent or not.
- 根据以上的模型和信息，最终识别一笔未知的交易是否为存在欺诈行为。



# Is this payment legitimate?

## 如何识别一笔支付纪录是合法的呢?

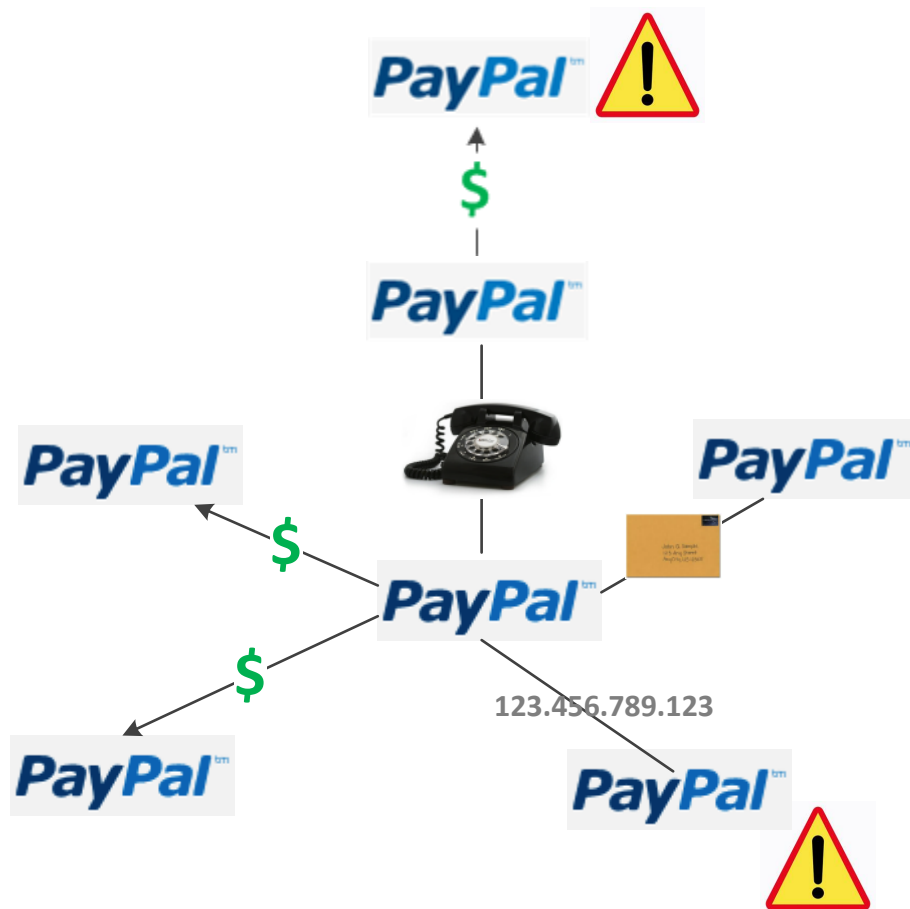
GOOD?



BAD?



THE JUDGMENT IS BASED ON...  
The transaction itself and its neighbors  
利用交易图中的每笔交易纪录及其邻居信息...



# TO START WITH... THE SIMPLEST CASE

Start with the simplest case, with only accounts and transactions information in the graph.

最简单例子〉交易图中只存在帐户信息及其交易纪录

For every fraudulent transaction (from historical data), build a network around it.

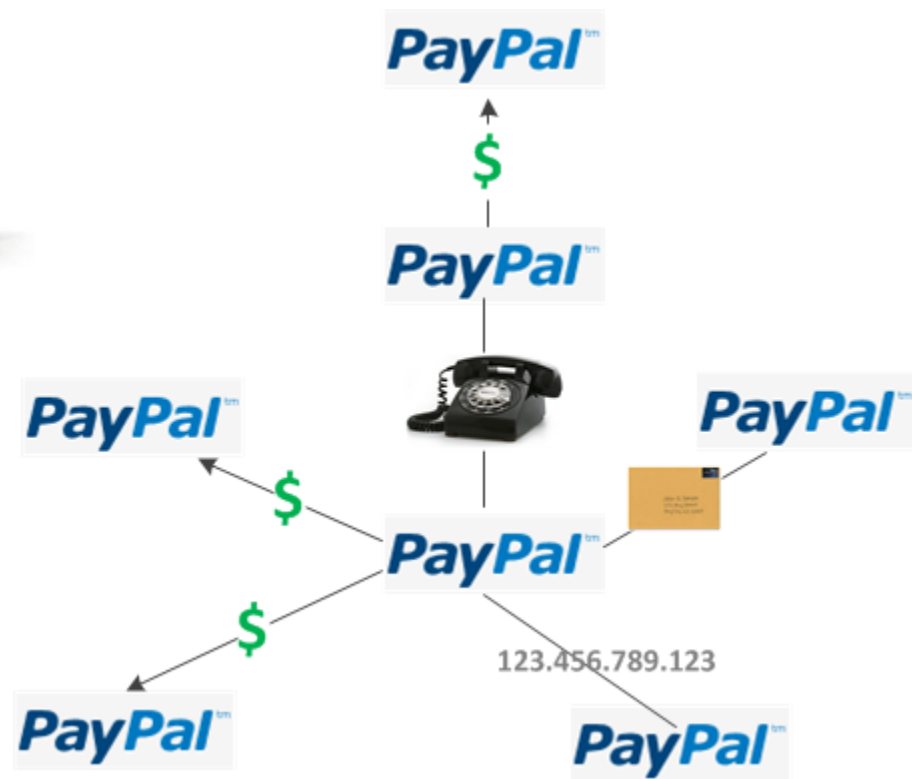
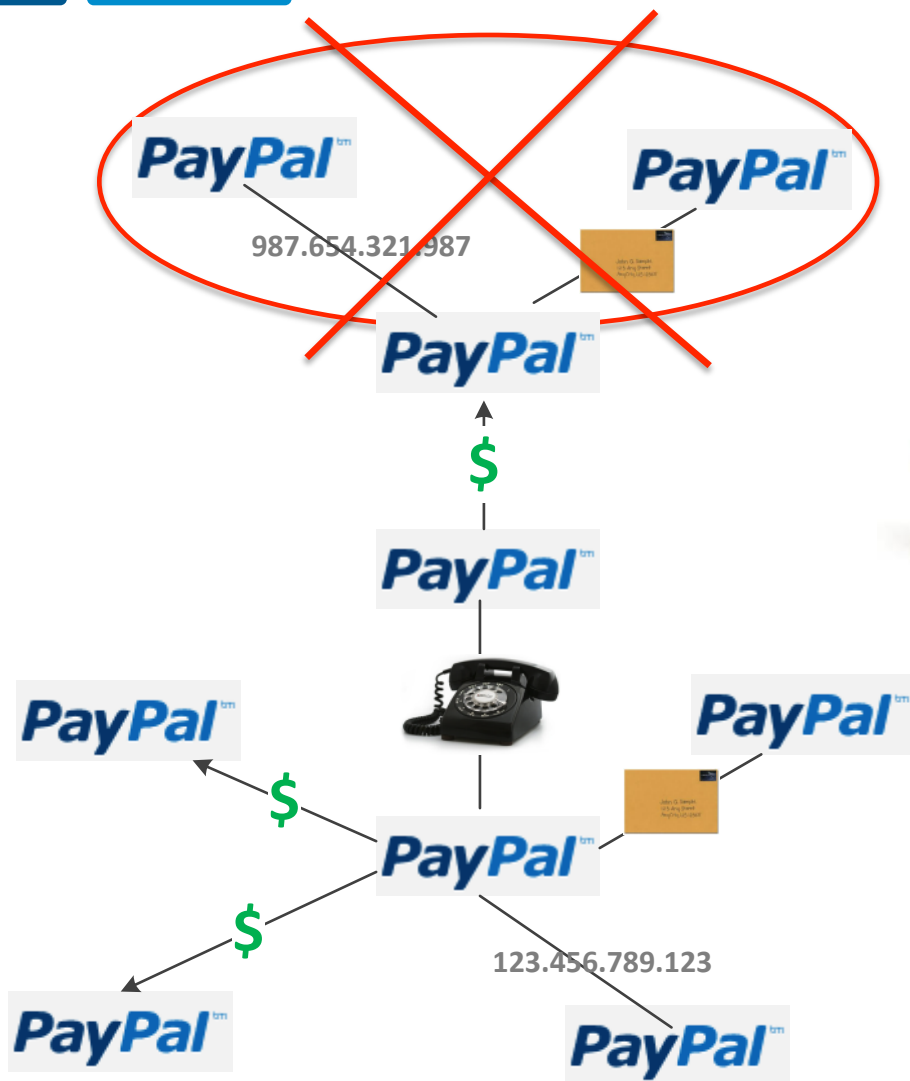
利用历史数据针对每笔欺诈交易纪录，构建该欺诈交易纪录对应的子图

Perform a clustering on them based on the graph similarity.

利用特定的图类似度量方式，构建聚类算法

# HOW DO YOU DEFINE SIMILARITY?

High Similarity



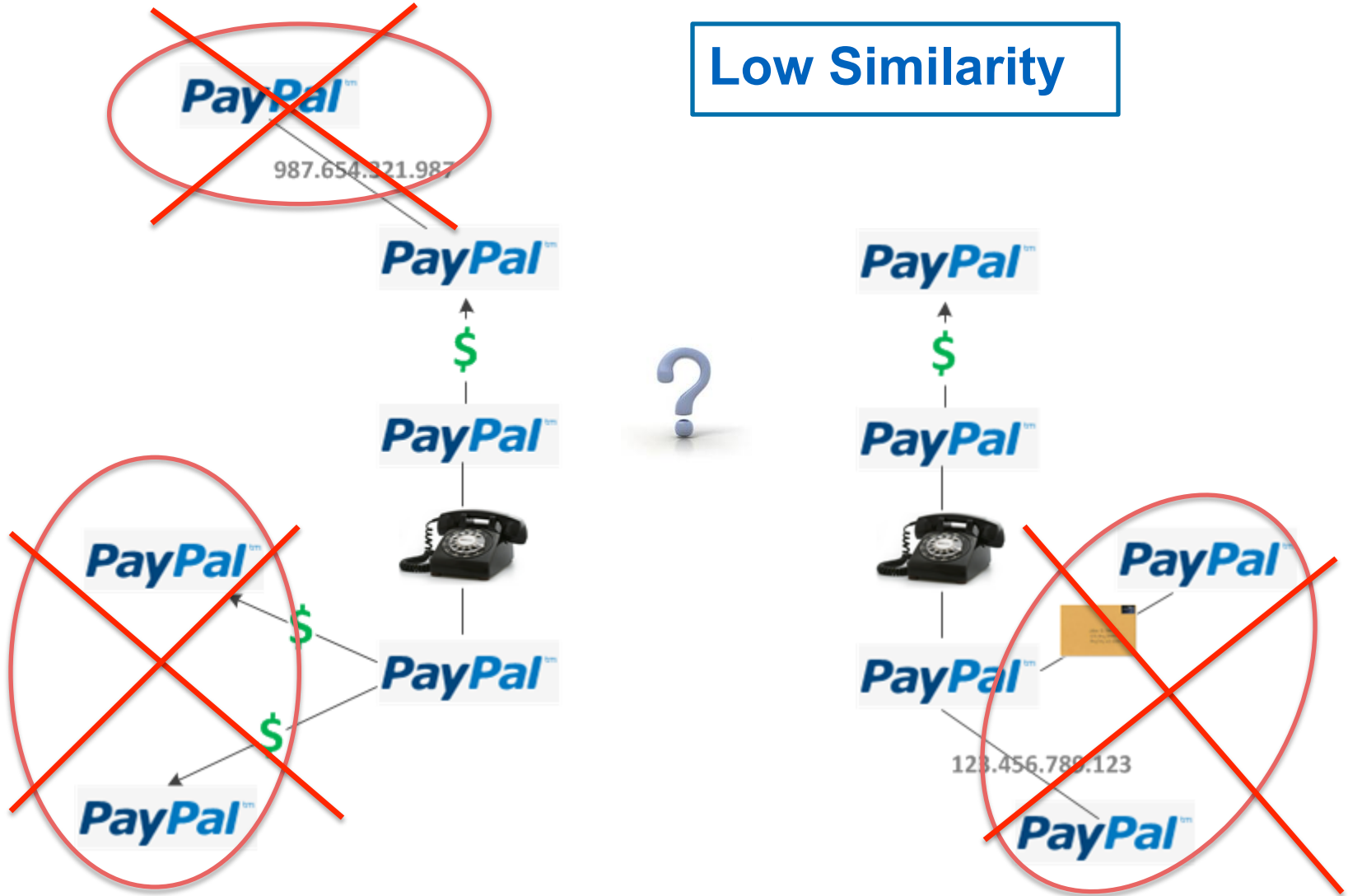
同濟大學  
TONGJI UNIVERSITY

PayPal™



# HOW DO YOU DEFINE SIMILARITY?

Low Similarity



同濟大學  
TONGJI UNIVERSITY

PayPal™

# TO SCALE UP... COMPLEX SITUATIONS

More information & relations should be taken into consideration (e.g. IP Address, Phone number...)

需要考虑更多的其他信息，包括IP地址，电话号码，email地址等


With increasing size of data, distributed system with index should be designed to store & process it.

当数据规模急剧增大之后，可扩展的分布式系统成为必然

How to meet the need of real-time processing?

难点是如何进行实时的处理

# RELEVANT TECHNIQUES



## Machine Learning

- Clustering (K-Means)
- Classification (Bayesian Network)

## Graph Mining

- BSP Model (Apache Hama)
- MapReduce

## Storage

- Hadoop(HDFS)
- Distributed Indexing
- Heterogenous Graph Model

# DATA STORAGE

## HADOOP(HDFS)



A framework that allows for the distributed storing and processing of large data sets across clusters of computers.

Built-in Fault-tolerance and synchronizing mechanism.

Designed to scale up from single servers to thousands of machines, each offering local computation and storage.

# DATA STORAGE INDEX



An index framework that allows to add index for the dataset based on the data schema.

Index framework also helps to improve the real time data processing and query

A separate index storage space for better performance and rebuilding

# DATA STORAGE GRAPH



A Heterogeneous graph model will be used to present different features of node and edge

通过异构图模型来表达节点和边的不同特征

The similarity of “ user activity” can be change to the similarity of heterogeneous graph

用户行为的类似性可以反映到异构图的类似性。

With the support of index framework, the heterogeneous graph can be easily scanned and rebuilt

通过构建图索引框架，可以非常容易的扫描和构建异构图。

# GRAPH MINING APACHE HAMA



The  
**Apache Hama Project**  
<http://hama.apache.org/>

A pure BSP (Bulk Synchronous Parallel) computing framework on top of HDFS.

HDFS之上的BSP计算框架

Designed for massive scientific computations such as matrix, graph and network algorithms.

用于大型科学计算，比如矩阵，图和网络算法

Supports message passing paradigm style of application development.

可以支持基于消息传递的应用开发

# APACHE MAHOUT



Machine learning and data mining library  
机器学习与数据挖掘的开发库

Implemented on top of Apache Hadoop using Map/Reduce paradigm

Apache Hadoop之上的实现

Scalable to reasonably large datasets  
可扩展到大数据集





# HOW DO WE CLUSTER GRAPHS?

## K-MEANS ALGORITHM

A rather simple but well known algorithms for grouping objects, clustering.

非常简单但是广为所用的聚类算法

All objects need to be represented as a set of numerical features.  
所有的数据对象均表达成以数值为特征的集合

The similarity is measured by the distance of those features.  
通过特征值之间的距离来测量数据对象之间的类似度

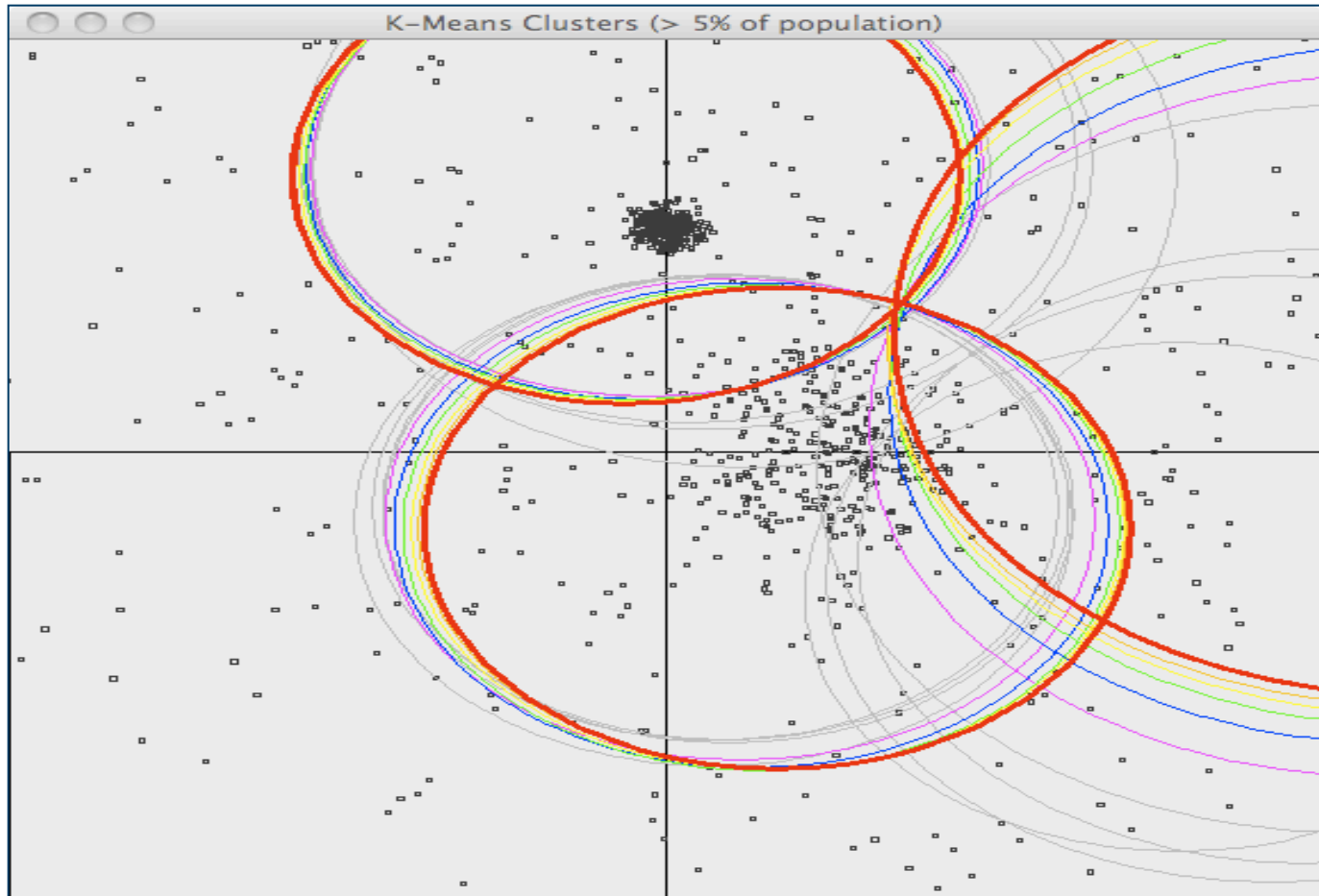
# HOW DO WE CLUSTER GRAPHS?

## SAMPLE FEATURES

- Number of Nodes
- Number of Edges
- Number of restricted accounts to all the accounts
- Number of different email\_domain
- Maximum value of the accounts' creation\_time
- ...

# HOW DO WE CLUSTER GRAPHS?

## SAMPLE OUTPUT



# HOW DO WE CLASSIFY GRAPHS?

## Naïve Bayesian朴素贝叶斯



Supervised learning algorithm



Popular classification algorithm



Assumptions:

- All attributes are equally important
- All attributes are statistically independent

# NAÏVE BAYESIAN MODEL

## Bayes Rule

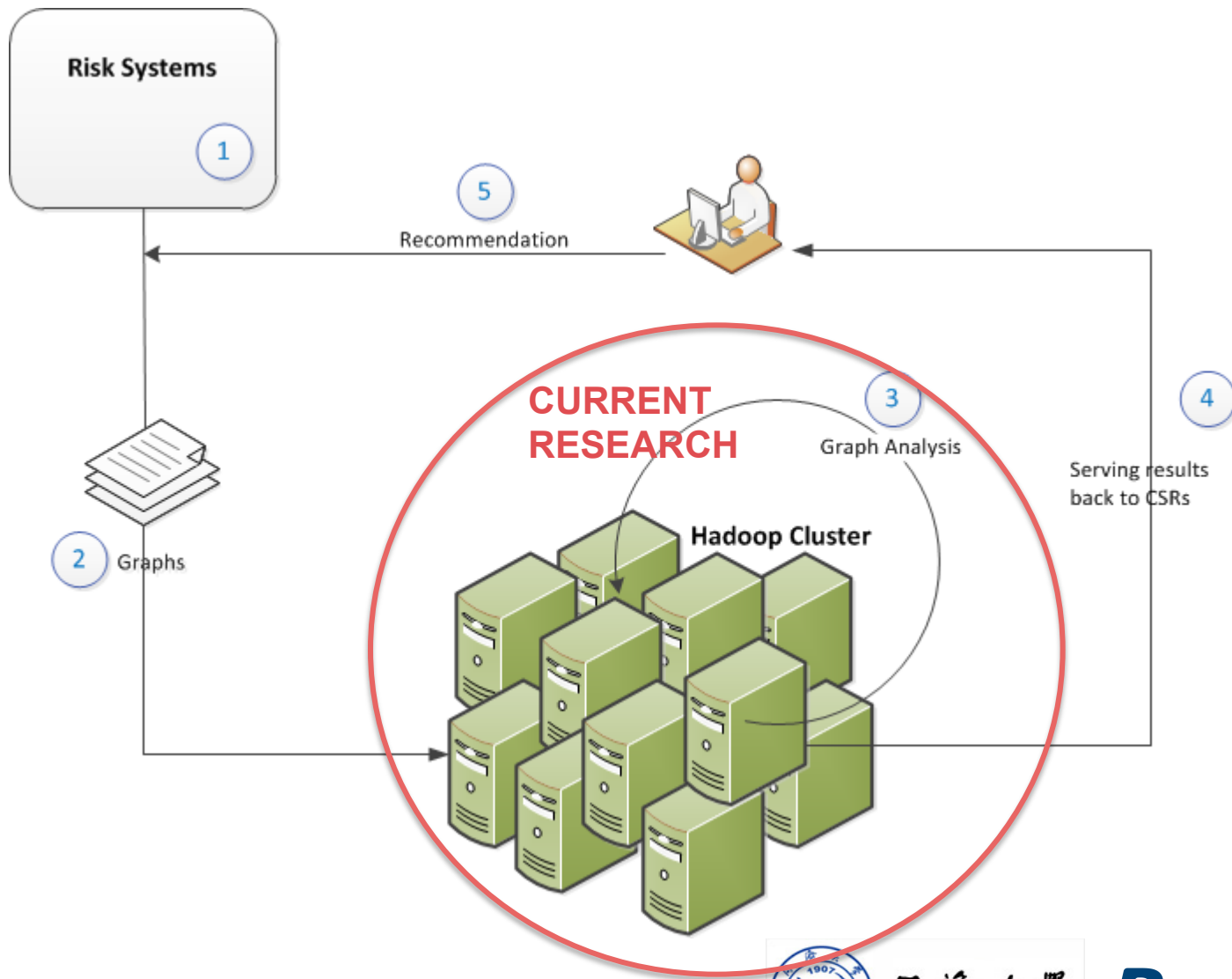
Posterior probability = Prior probability \* likelihood

$$P(C_i | f_1, f_2, \dots, f_n) = P(C_i) * P(f_1, f_2, \dots, f_n | C_i)$$

Where,

- $P(C_i | f_1, f_2, \dots, f_n)$  = probability of a feature set  $F$  is in category  $C_i$
- $P(C_i)$  = probability of a given category

# LONG TERM PICTURE



# Research Works in Academic Communities

## 分布式图的研究问题

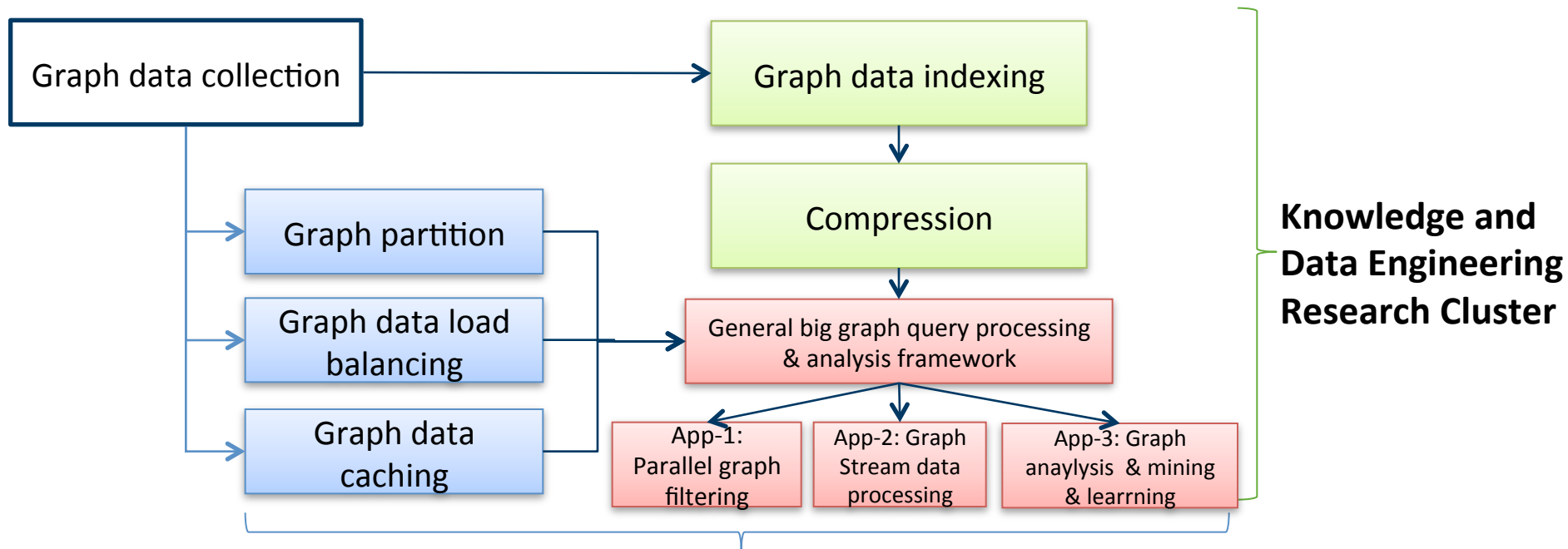
### Framework:

1. CMU pegasus (KDD'11, ICDM'09 best app paper)
2. CMU PowerGraph, GraphChi (OSDI'12), GraphLab (VLDB'11)
3. MSRA **Trinity** (SIGMOD'13)

### Key problem:

1. Graph partition: SPAR (SIGCOMM'10), Sedge (SIGMOD'13),
2. Streaming graph partition (KDD'12)
3. Subgraph query and filtering

# On going works 正在开展的工作



**High Performance and Distributed Computing Research Cluster**

24





THANK YOU

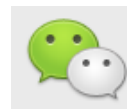


同濟大學  
TONGJI UNIVERSITY

**PayPal**<sup>™</sup>



@InfoQ



infoqchina

软件  
正在改变世界!