Pattern Recognition and Anomaly Detection in Bookkeeping Data

Asian Bureau of Finance and Economic Research (ABFER)

8th Annual Conference – Accounting | May-26, 2021

By Pierre J. Liang

Carnegie Mellon University

Coauthors Influences

- Aluna Wang, Tepper School of Business at CMU
- Christos Faloutsos, School of Computer Science at CMU
- Leman Akoglu, Heinz School and School of Computer Science at CMU

Outline

Use Case 1: Use Graph Mining Tools to Detect Journal entry anomalies

Broad Lessons

Use Case 2: Use Graph Mining Tools to Quantify Account Classification and Aggregation

Why this is cool?

USE CASE:

Using Graph Mining to Detect Anomalies in Bookkeeping Data

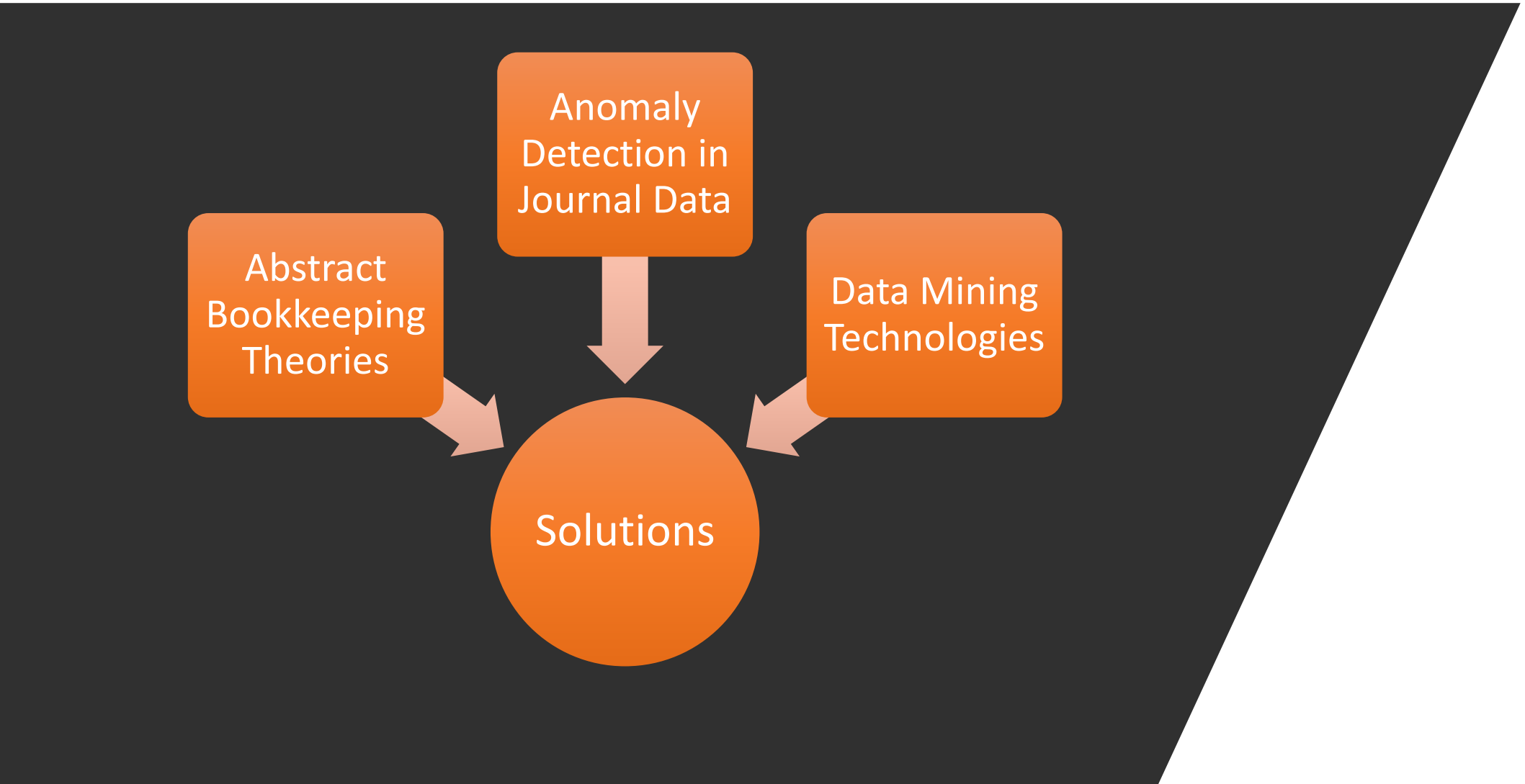
Principal Investigators:

Leman Akoglu and Pierre J. Liang

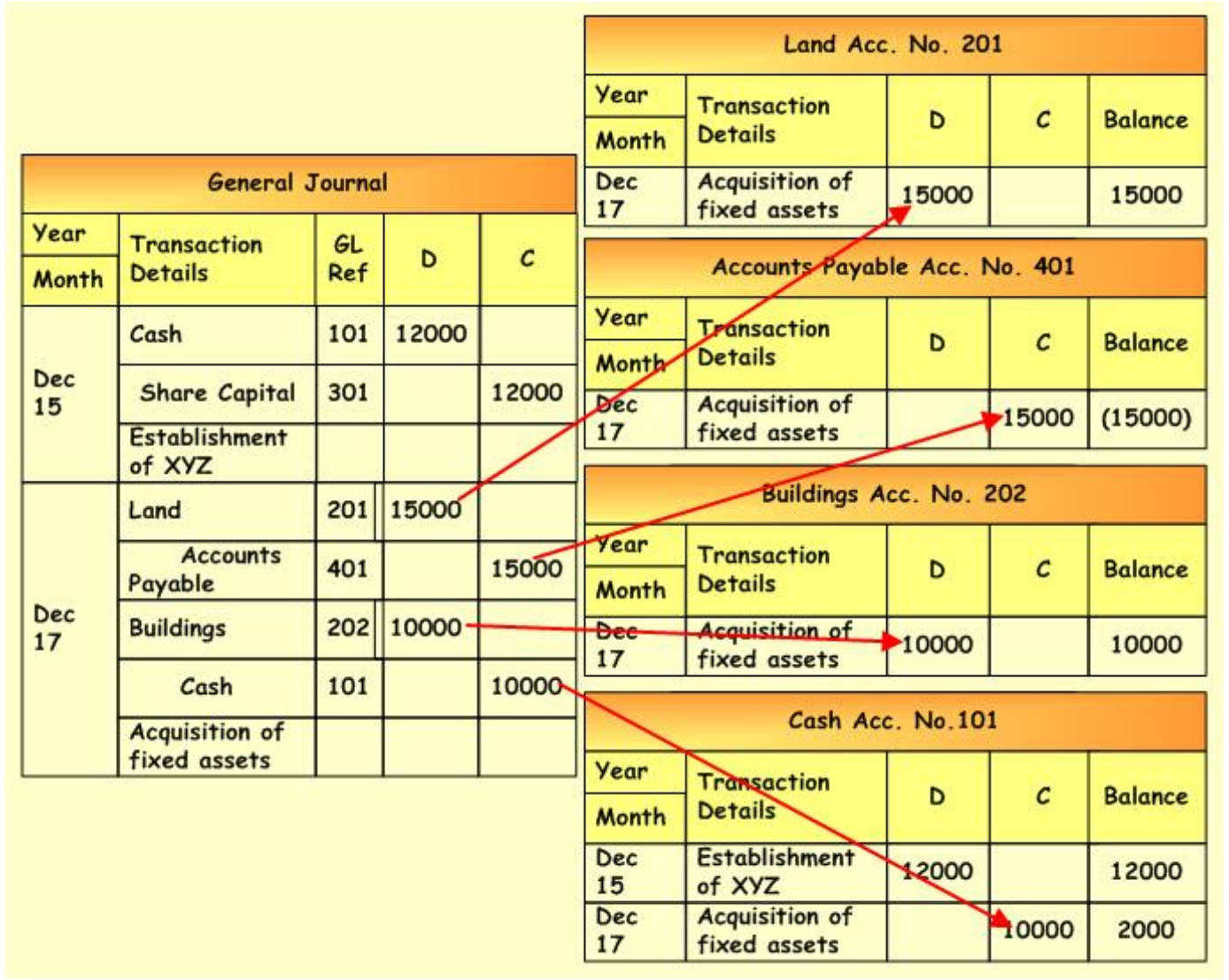
Carnegie Mellon University

Three Ingredients

Ingredient 1: Data and Problems from Field



Journals and Ledgers



General Journal

Year / Month	Transaction Details	GL Ref	D	C
Dec 15	Cash	101	12000	
	Share Capital	301		12000
	Establishment of XYZ			
Dec 17	Land	201	15000	
	Accounts Payable	401		15000
	Buildings	202	10000	
	Cash	101		10000
	Acquisition of fixed assets			

Land Acc. No. 201

Year / Month	Transaction Details	D	C	Balance
Dec 17	Acquisition of fixed assets	15000		15000

Accounts Payable Acc. No. 401

Year / Month	Transaction Details	D	C	Balance
Dec 17	Acquisition of fixed assets		15000	(15000)

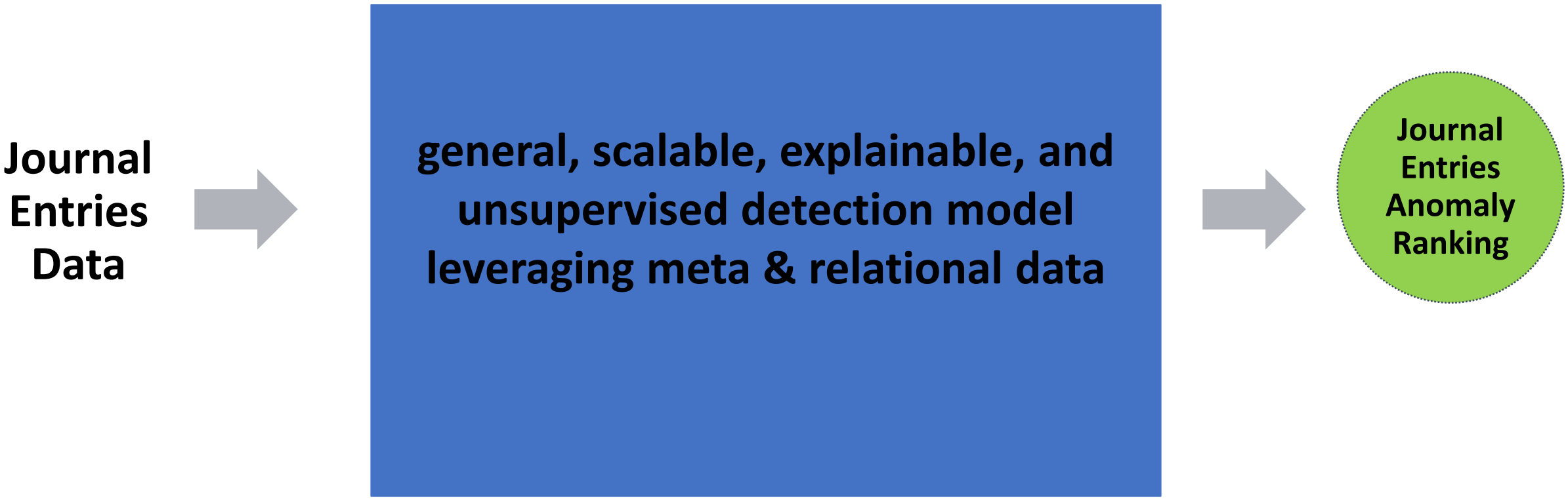
Buildings Acc. No. 202

Year / Month	Transaction Details	D	C	Balance
Dec 17	Acquisition of fixed assets	10000		10000

Cash Acc. No.101

Year / Month	Transaction Details	D	C	Balance
Dec 15	Establishment of XYZ	12000		12000
Dec 17	Acquisition of fixed assets		10000	2000

Our Project GOAL



Solutions

General Journal Data

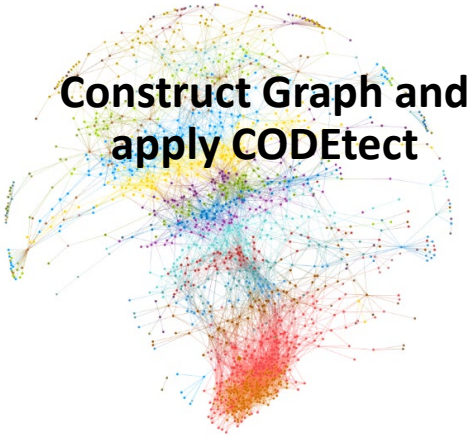
Data cleaning & filtering

Meta-data

Apply CompreX to Meta-data

Anomaly ranking based on the CompreX scoring system

Relational data

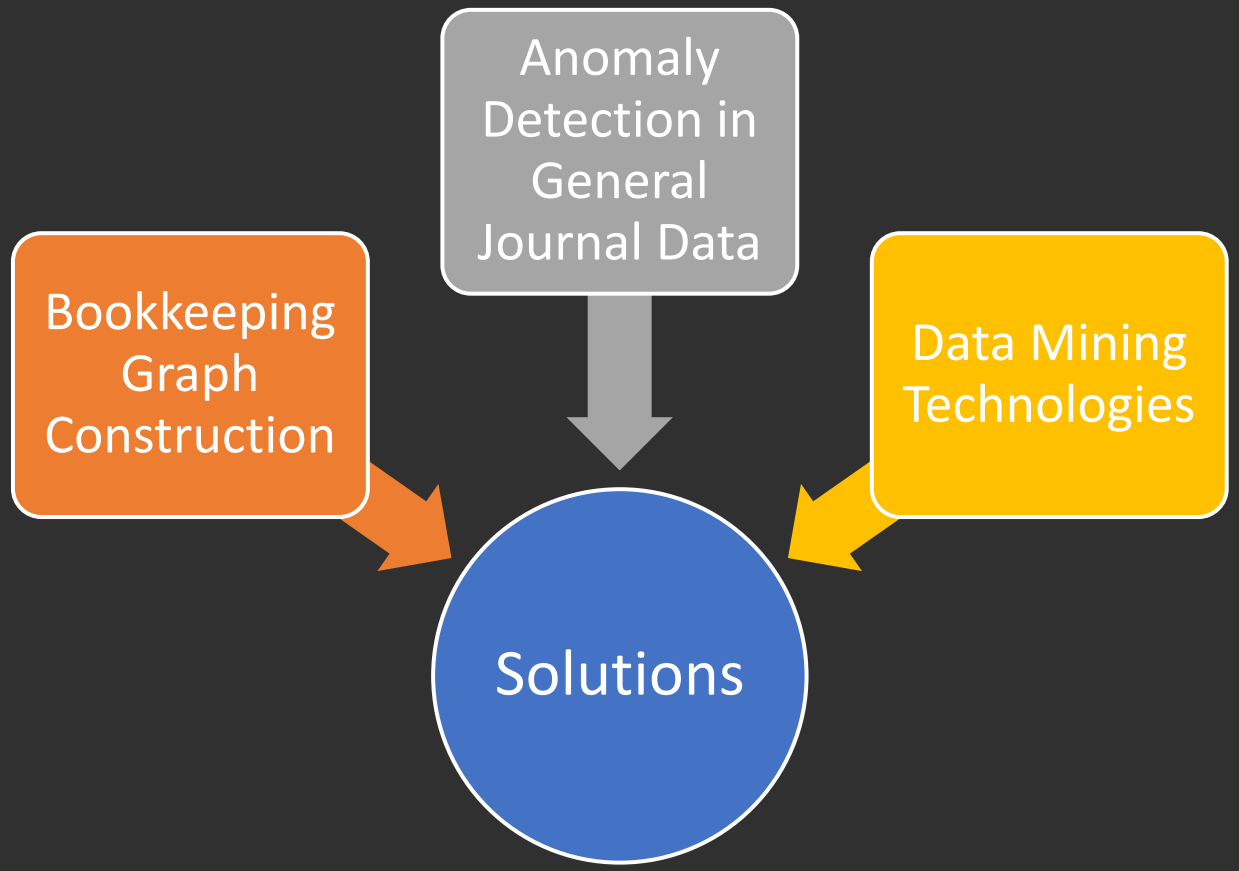


Construct Graph and apply CODEtect

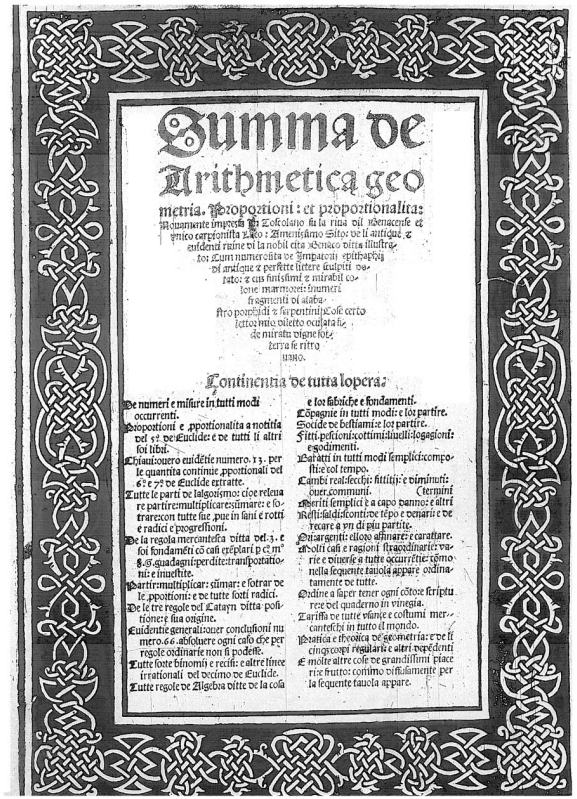
Anomaly ranking based on the CODEtect scoring system

The Combined Ranking

Ingredient 2: Abstract Bookkeeping Theories



Luca Pacioli: Summa de Arithmetica … (1494 AD)

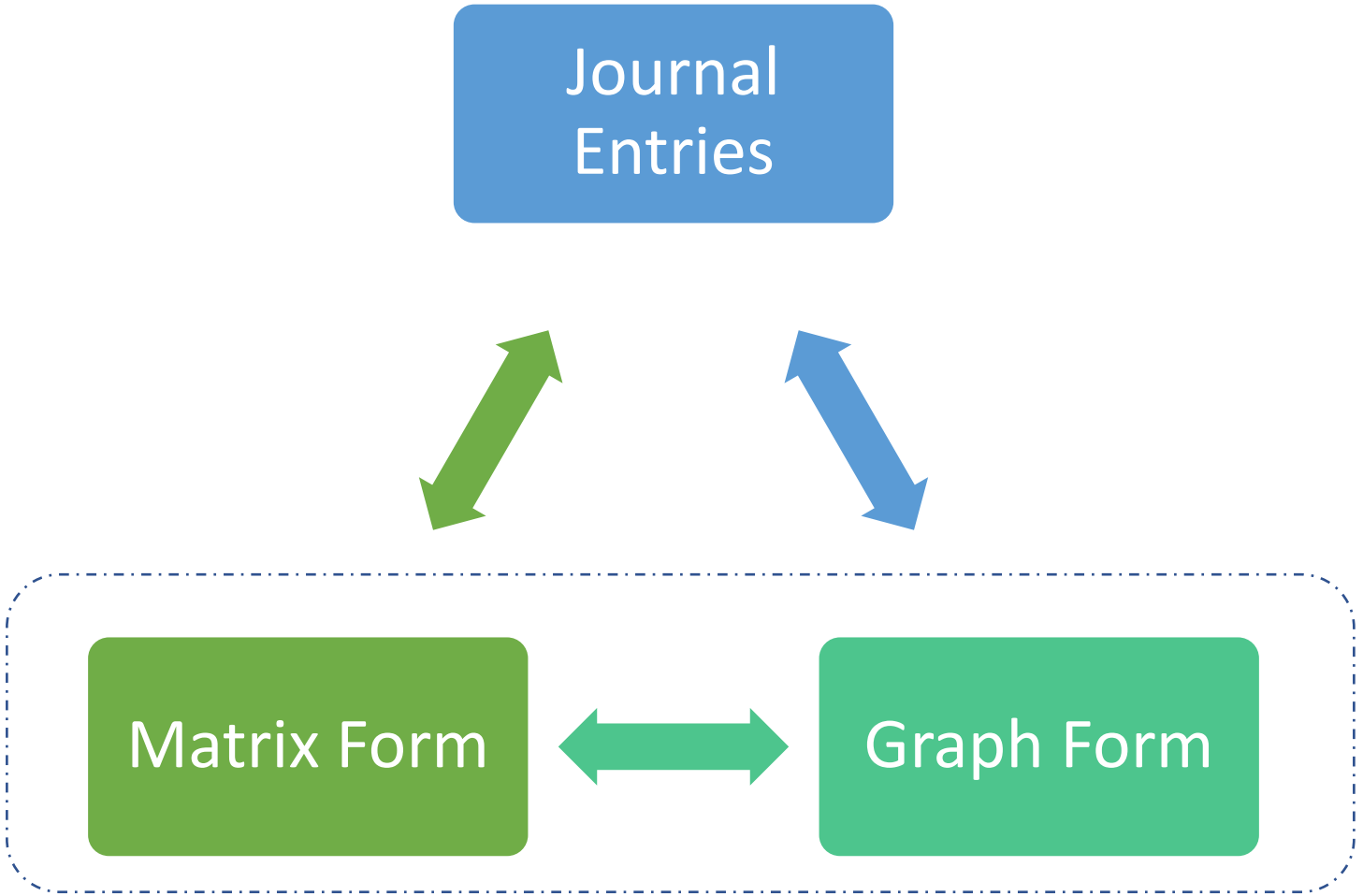


Summa de
Arithmetica geo
metria. Proportioni : et proportionalita:
Nouamente impressa per Paciolano fra la ripa dil Benacense et
pnico carpentista Luca : Almenisamo Sito: de li antiqui. et
eusdenti ruine di la nobil cita Benaco ditta illustra-
tor. Cum numerosita de Epitaphij epithaphij
di antiqui et perfette littere sculpiti de-
tato: et cus finissimi et mirabil co-
lone marmoree: numeri
fragmenti di alaba-
stro pontidi: et sarpentini: cose certo
latori mira: oltreto ouiata fe.
de miranda vigne fot.
terra se ritro
uano.

Continentia de tutta lopera.

De numeri e misure in tutti modi occurrenti.
Proportioni e ppotionalita a notitia vel 5?. de Euclide: e de tutti li altri sol libri.
Chiamouero e euidētie numero. 13. per le quantita continie ppotionali del 6?. e 7?. de Euclide extratte.
Tutte le parti de lalgoritmo: cioe releua re partire: multiplicare: simare: e sotrare: con tutte sue ppe in fani e rotti e radici e progressioni.
De la regola mercantesca ditta del. 3. e sol fondamēti cō casi ēpētari p cē m° §. §. guadagni: perdite: trasportatio ni e inuestite.
Partir: multiplicar: simar: e sotrar de le ppotioni: e de tutte sorti radici.
De le tre regole del Catayn ditta posi tione: e sua origine.
Euidentie generali: cioe conclusioni nu mero. 66. absolutere ogni caso che per regole ordinarie non si podesse.
Tutte sorte binomi e recisi: e altre linee irrationali del decimo de Euclide.
Tutte regole de Algebra ditte de la cosa e toe fabriche e fondamenti.
Cōpagnie in tutti modi: e toe partire.
Socide de bestiame: toe partire.
Sitti: positioni: cottimi: inuestiti: logagioni: e pdimenti.
Baratti in tutti modi semplici: compo stie col tempo.
Cambi reali: secchi: fittitij: e viminuti: pue: communi. Cōtrami
Meritij semplici: e a capo danno: e altri
Restisaldi: sconti: de tēpo e denari: e de recare a vn di piu partite.
Oro: argenti: elloro simare: e carattare.
Molti casi e ragioni straordinarie: vo rie e vtilissime: e tutte occurrētie: cō mo nella sequente tauola appare ordina tamente de tutte.
Ordine a saper tener ogni cōto: e scriptu ree: del quaderno: in venesia.
Tarifa: de tutte varie: e costumi: mer cantesche: in tutto il mondo.
Geratica: e theorica de geometria: e de li cinque corpi regulari: e altri: e pedenti
E molte altre cose de grandissimi piace rie: tutte: como: difficilmente per la sequente tauola appare.

Double Entry Bookkeeping's "Trifecta"

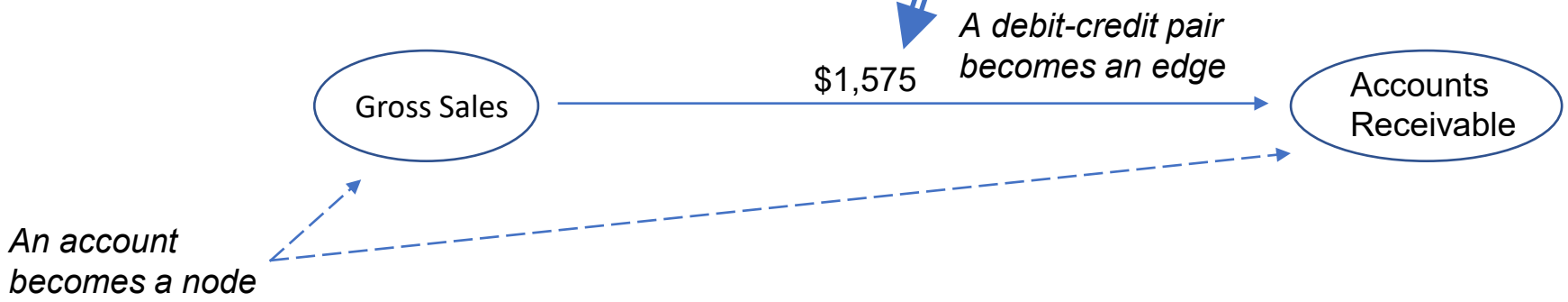


"Double-entry bookkeeping, especially its mathematical structure, has been the author's life-long mystery."
– Yuji Ijiri, 1993

Journal entries can be converted into a DiGraph

- Account => Node
- Entry => Edge
- Debit => Inflow
- Credit => outflow

GL_Account_Number	GL_Journal_ID	GL_User_ID	GL_Reporting_Amount	CA_FS_Caption	…
40060000 (Revenue)	2016/01-20160101 - 638621	JJSHEN	-1575	Gross Sales (GSL)	…
10415000 (Assets)	2016/01-20160101 - 638621	JJSHEN	1575	Accounts Receivable (ARV)	…
…	…	…	…	…	…

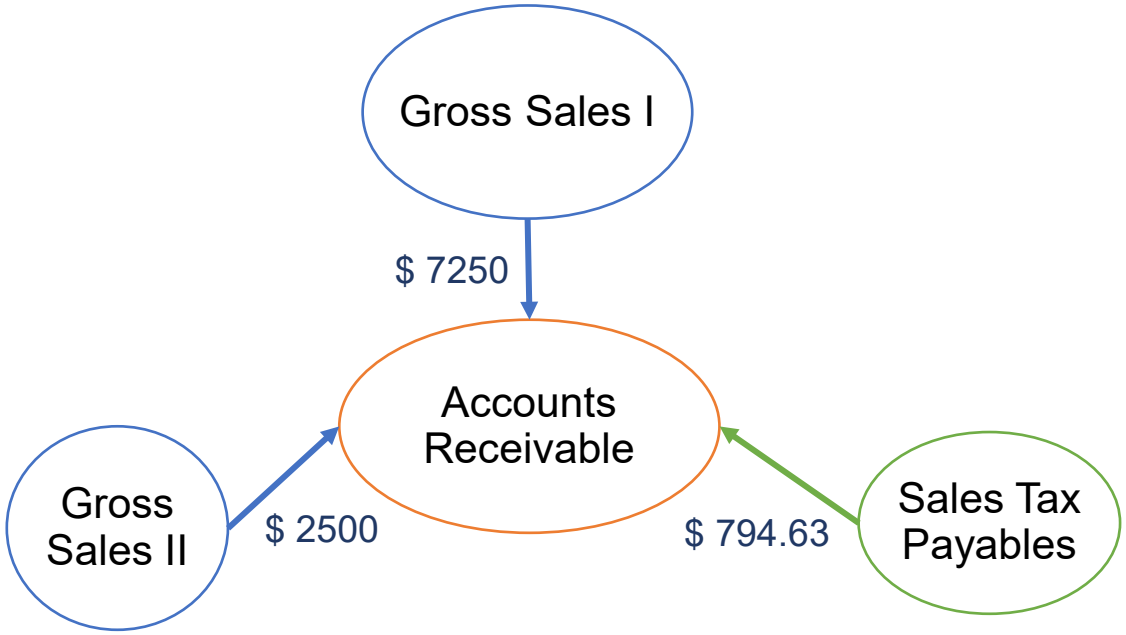


Journal entries can be converted into a DiGraph

HW 2016/01-20160101 -639167

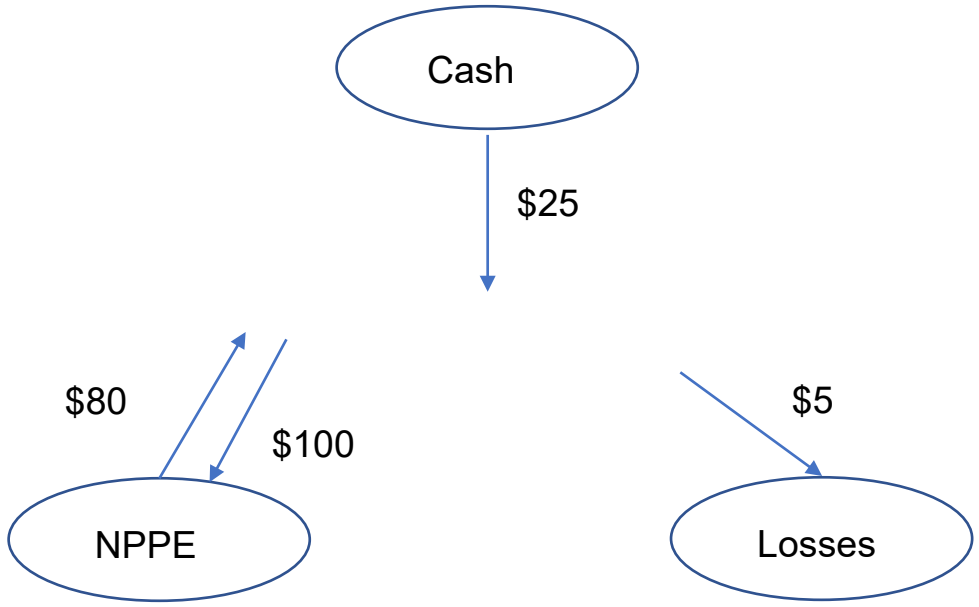
GL_Account_Number	CA_FS_Caption	Cr/Dr	GL_Reporting_Amount
40020000 (Revenue)	Gross Sales (GSL)	C	-7250
40020000 (Revenue)	Gross Sales (GSL)	C	-2500
20830000 (Liabilities)	Sales Tax Payables (STP)	C	-794.63
10390000 (Assets)	Accounts Receivable (ARV)	D	10544.63

Graph Construction Solution

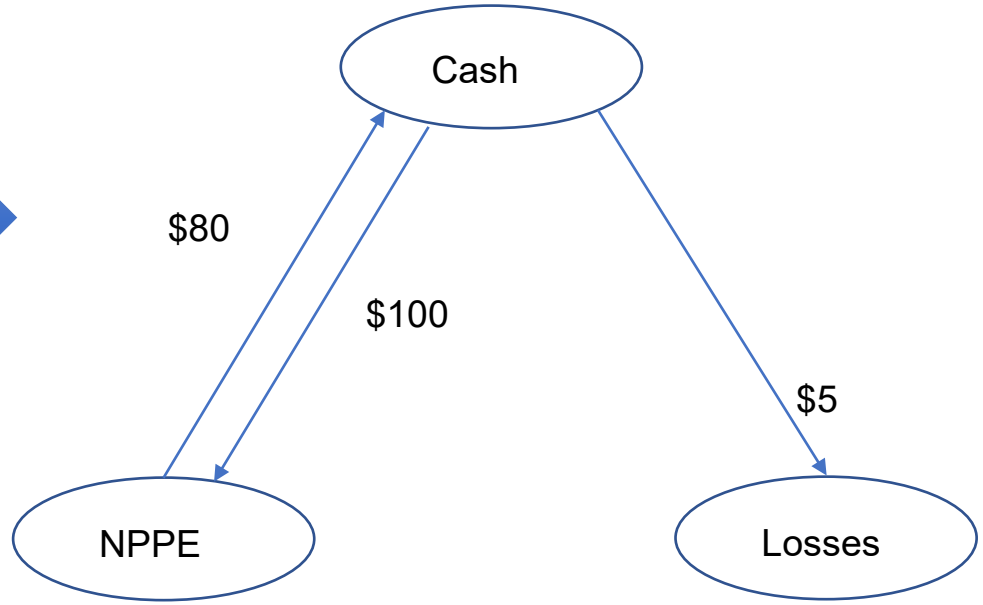


Journal entries can be converted into a DiGraph

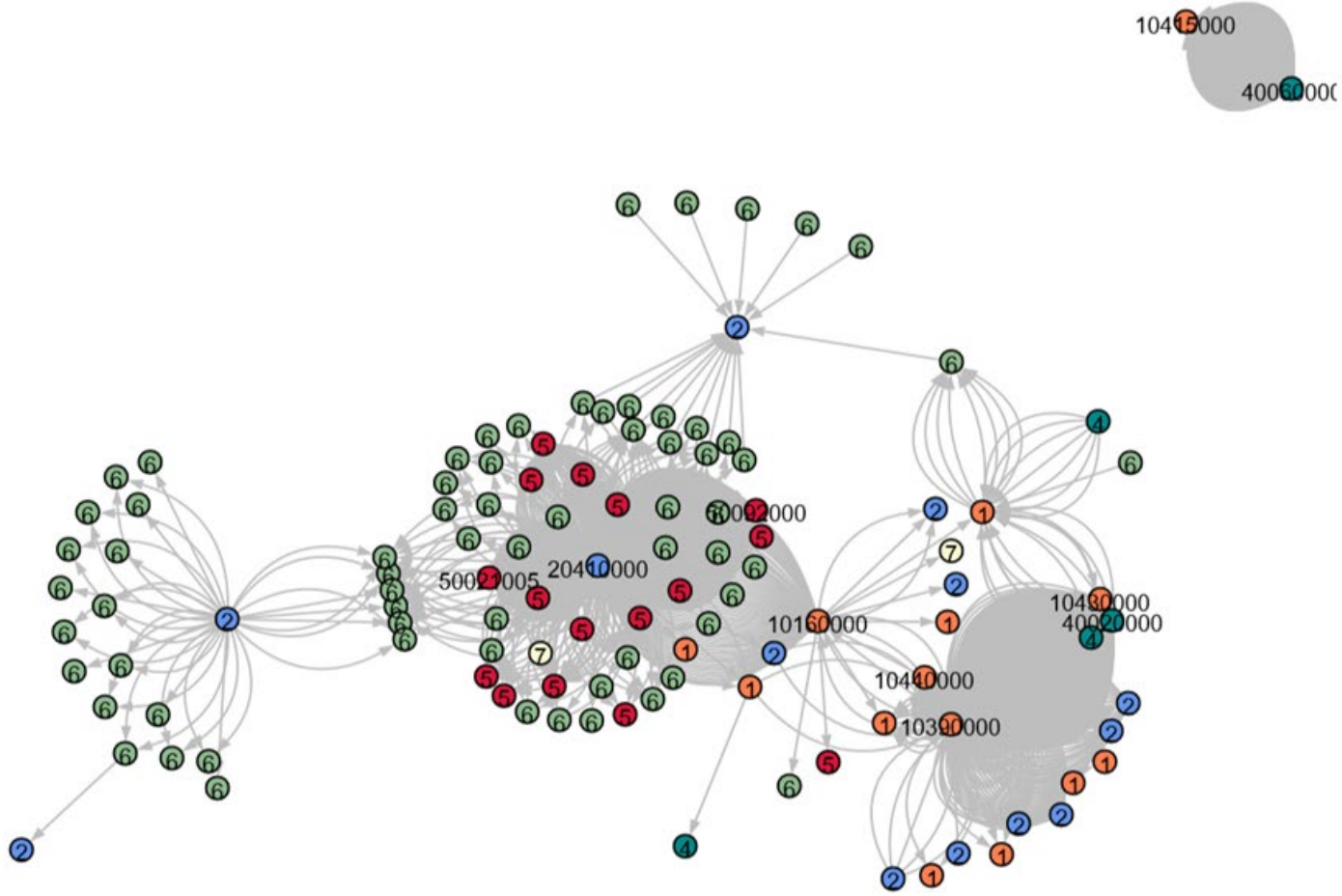
Compound Journal Entries



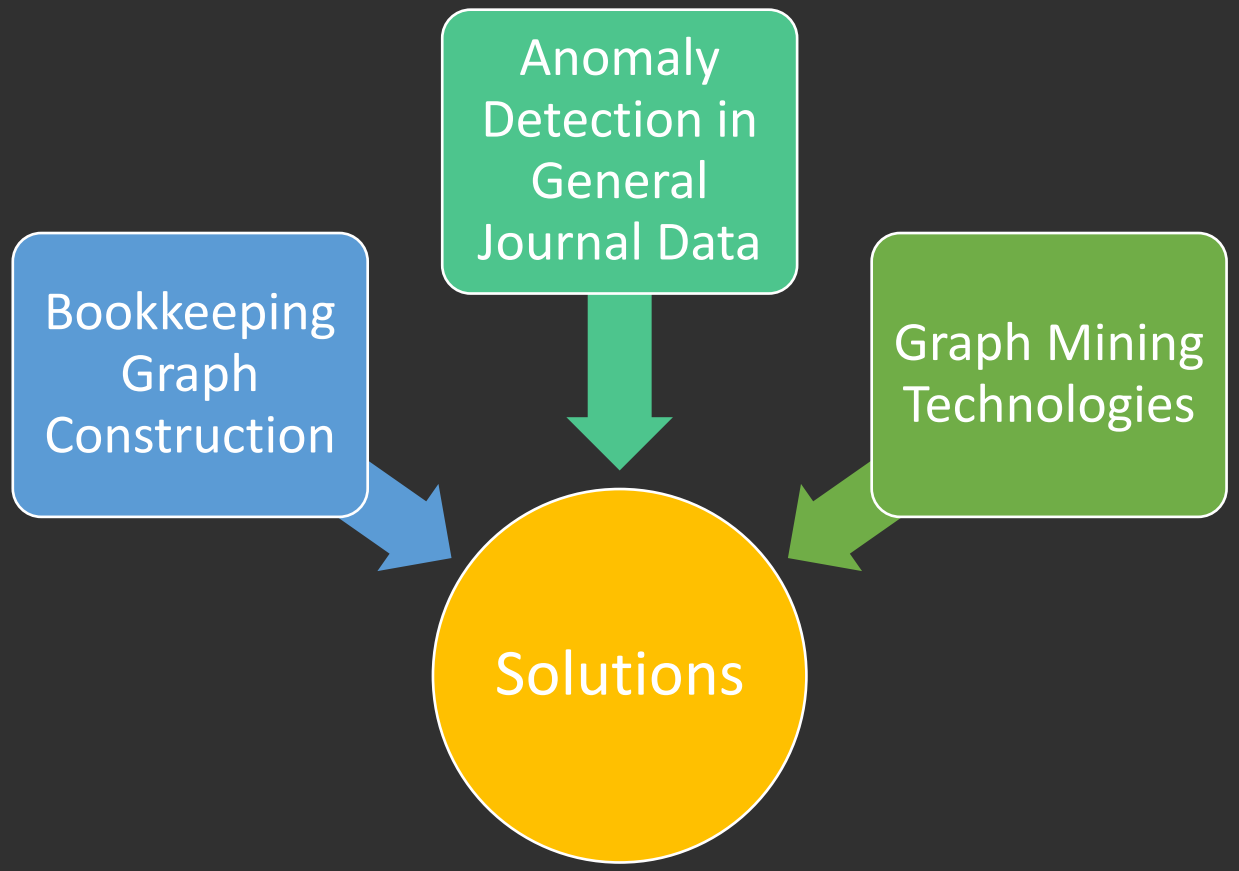
Graph Construction Solution



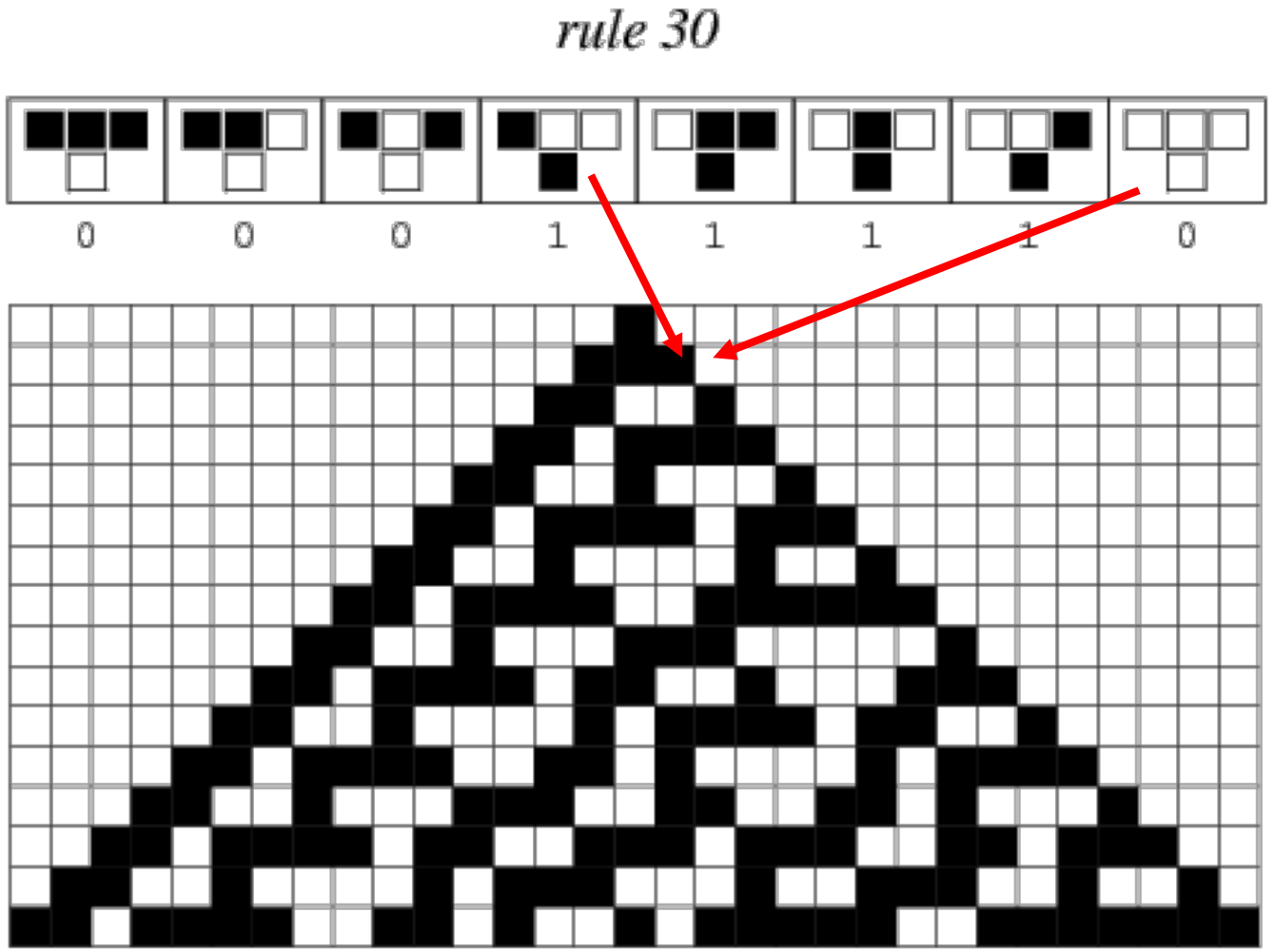
Bookkeeping Graph for a 10-day Window



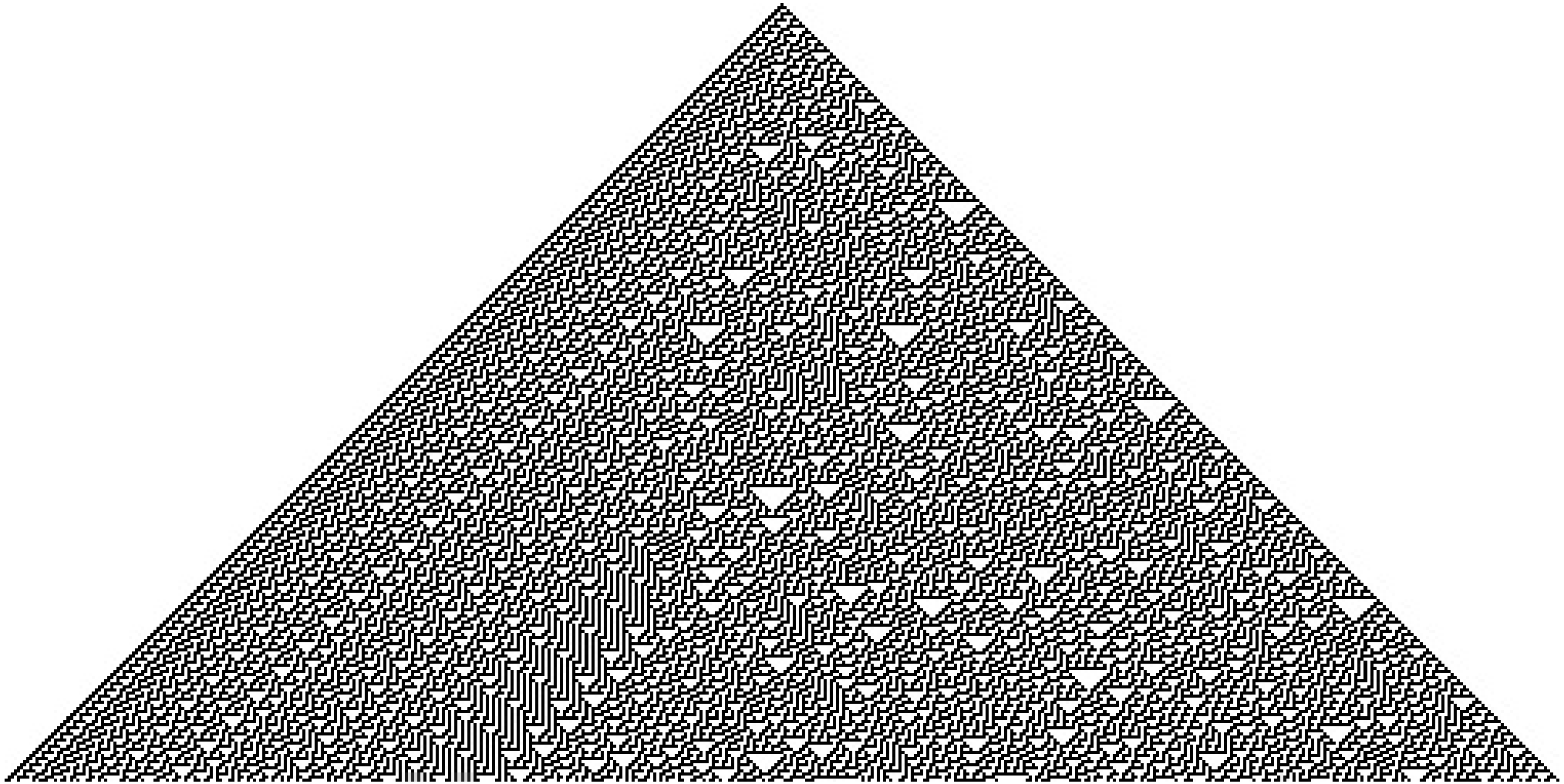
Ingredient 3: Data Mining Technologies



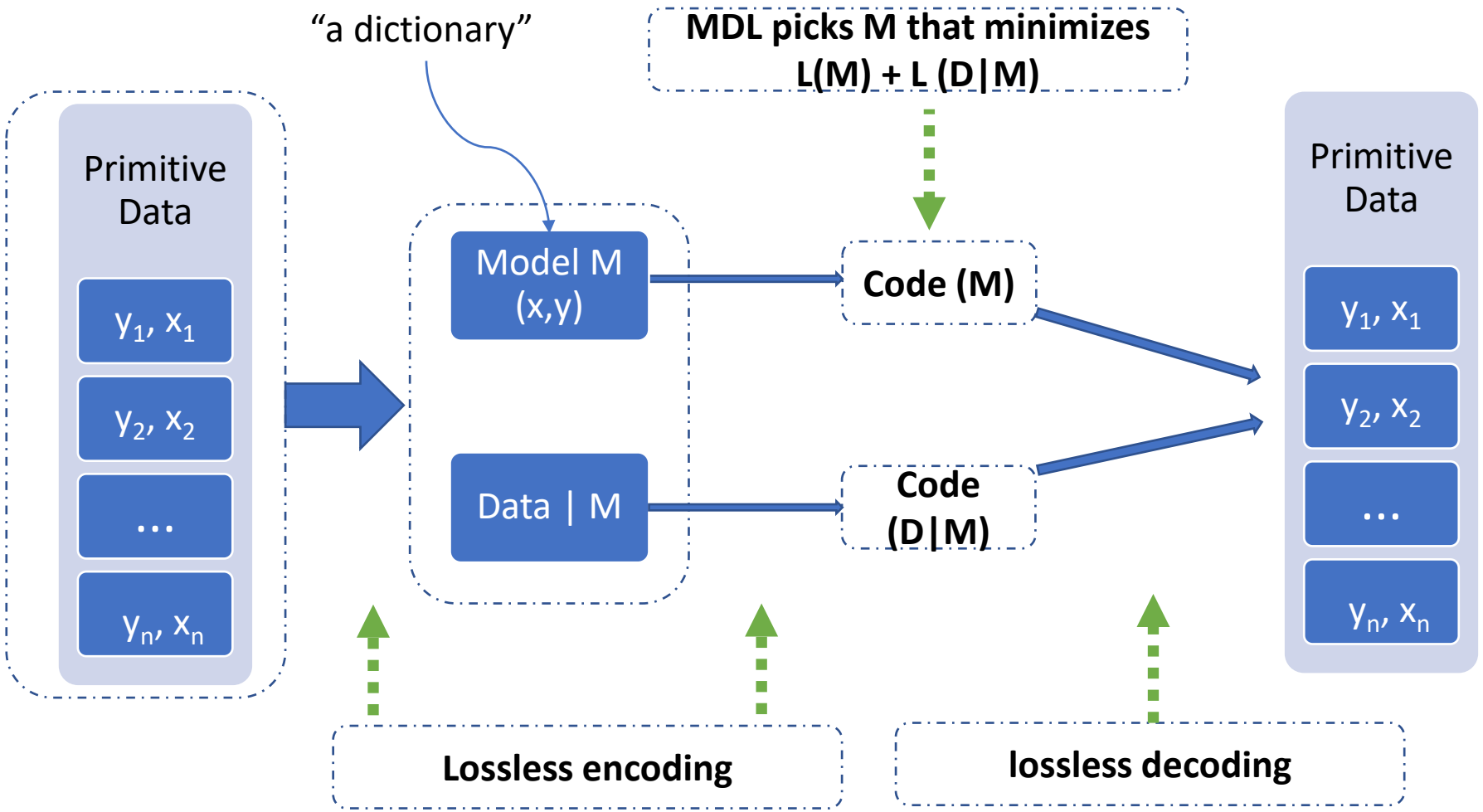
Stephen Wolfram's ECA Rule-30



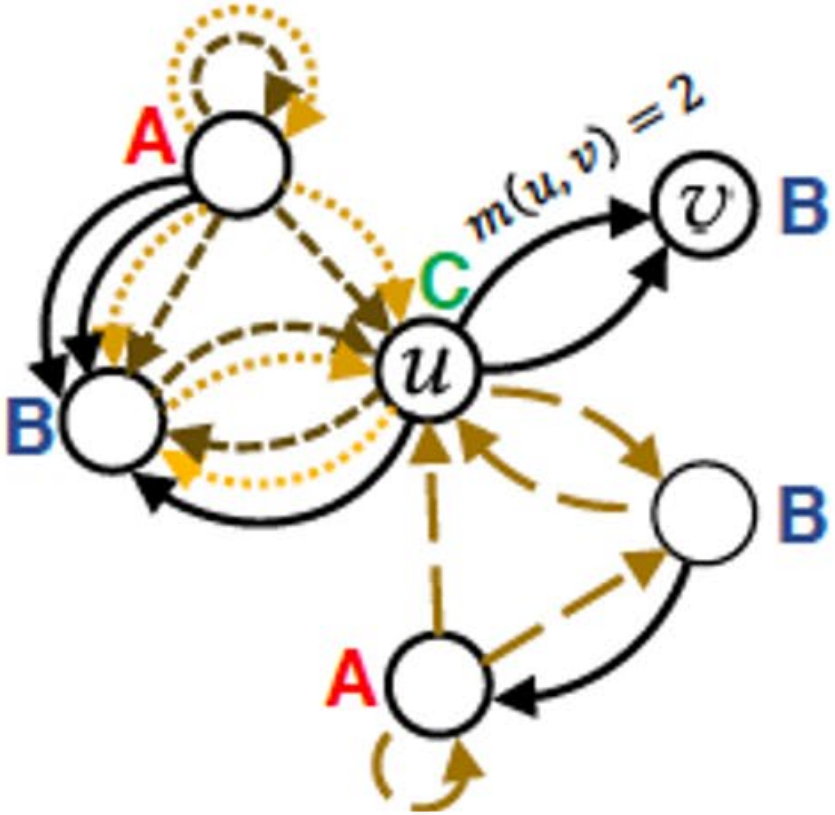
Stephen Wolfram's ECA Rule-30



Minimum Description Length (MDL) Principle: Compress data to recover patterns/models

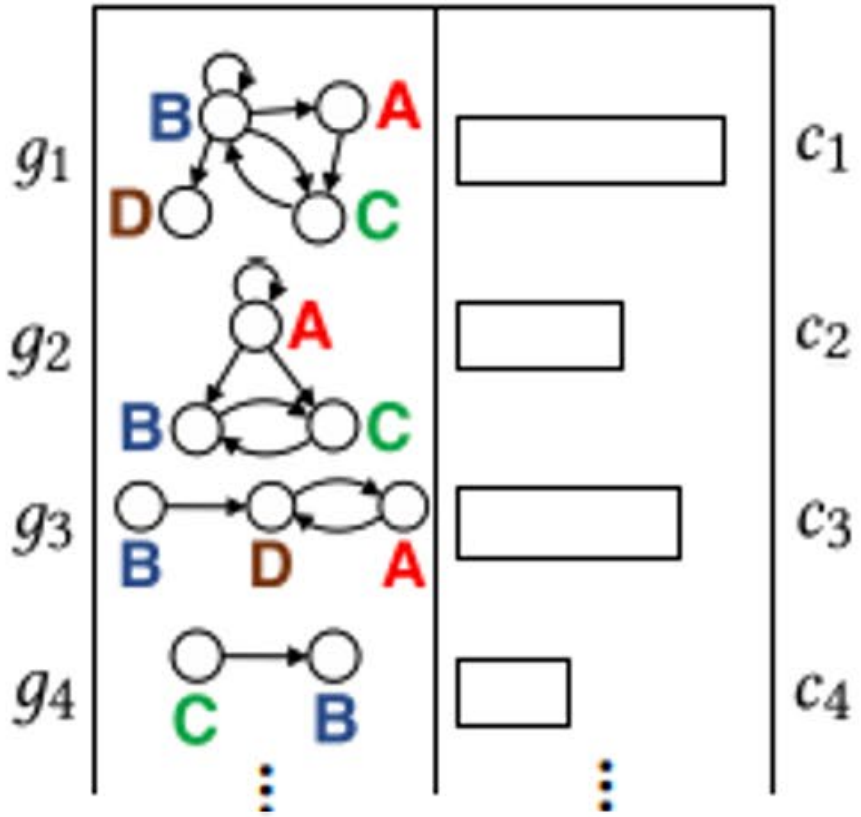


One transaction graph



(a)

Motif Table (MT)

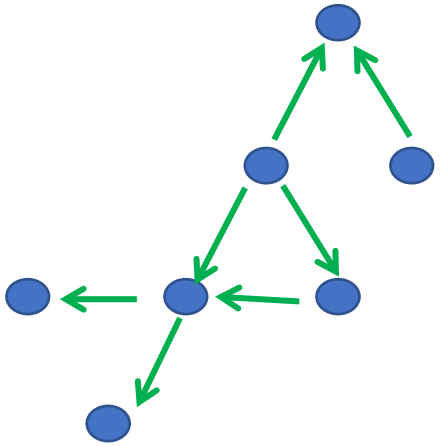


(b)

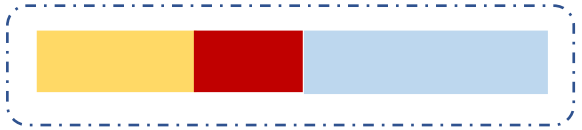
Graph Encoding via MDL Principle

A Transaction Journal Entry

=



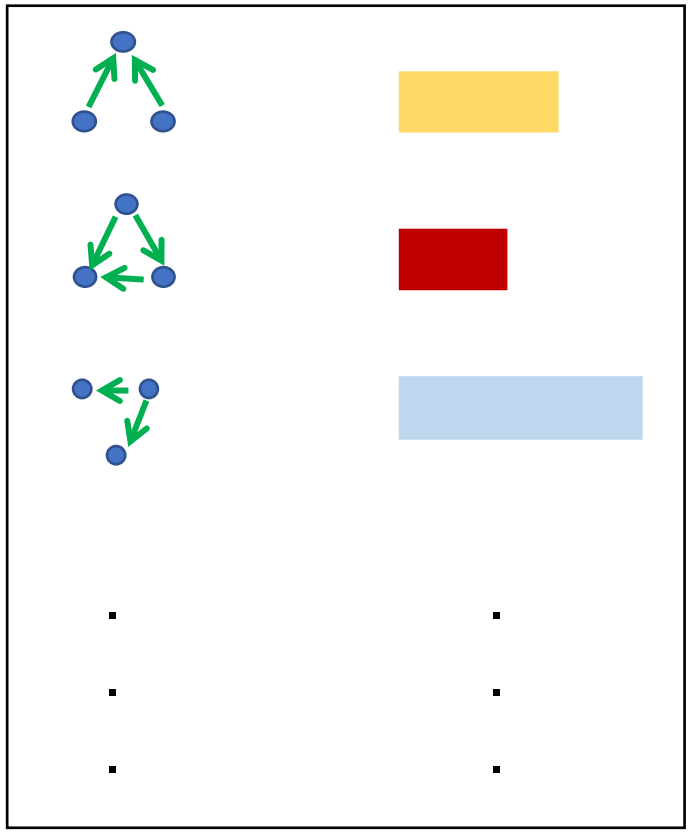
=



$L(G_i|MT)$

is the code length

Motif Table (MT)



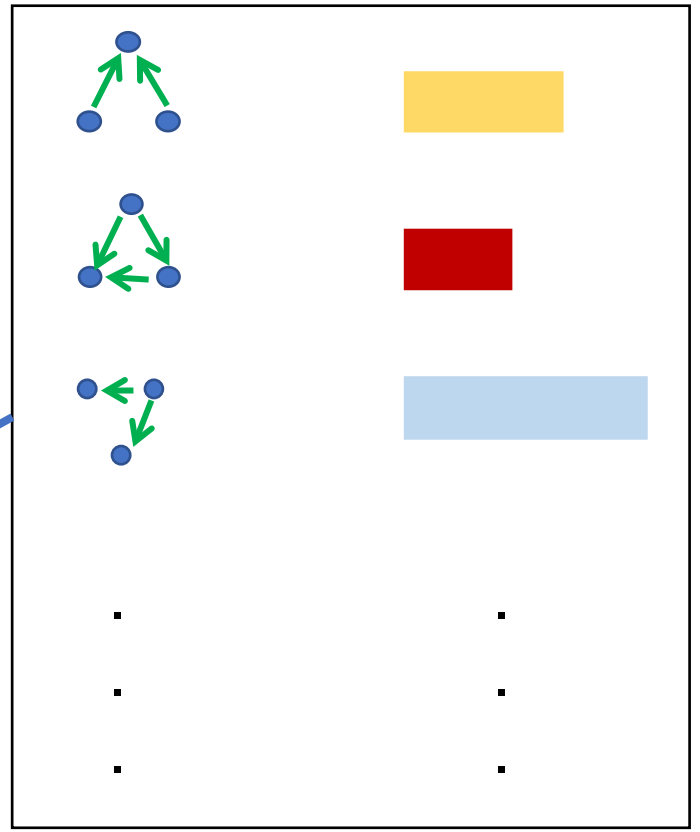
Graph Encoding via MDL Principle

$$\mathcal{G} = G_1 \; G_2 \; G_3 \; \ldots\ldots\ldots \; G_J$$

$$\sum_{i} L(G_i|MT) + L(MT)$$

Minimize by choosing a best Motif Table

Motif Table (MT)



Use Graph Encoding to ID Anomalies

G_1

5 Mofit_1 | 2 Mofit_25 | 1 Mofit_4

G_2

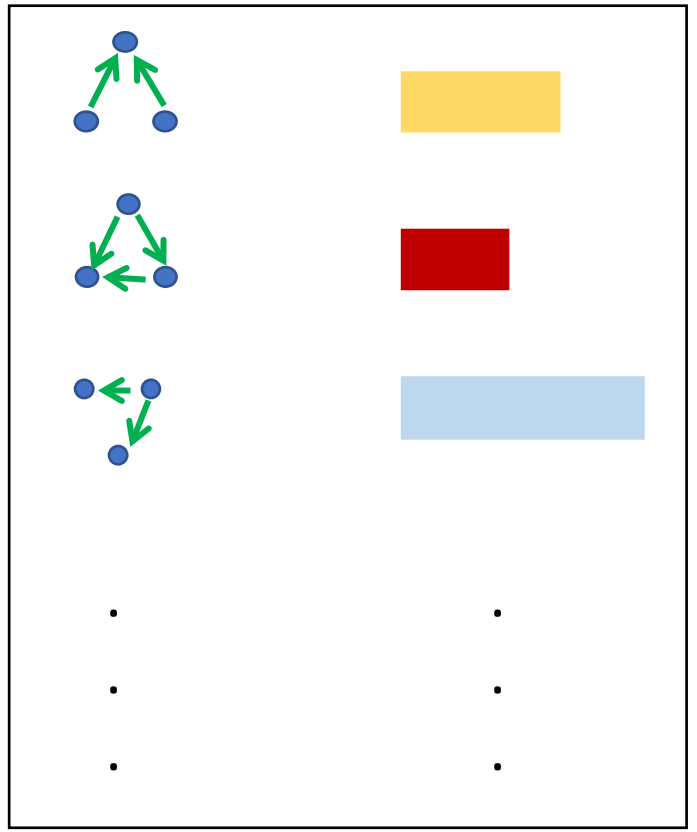
3 Mofit_1 | 2 Mofit_2

....

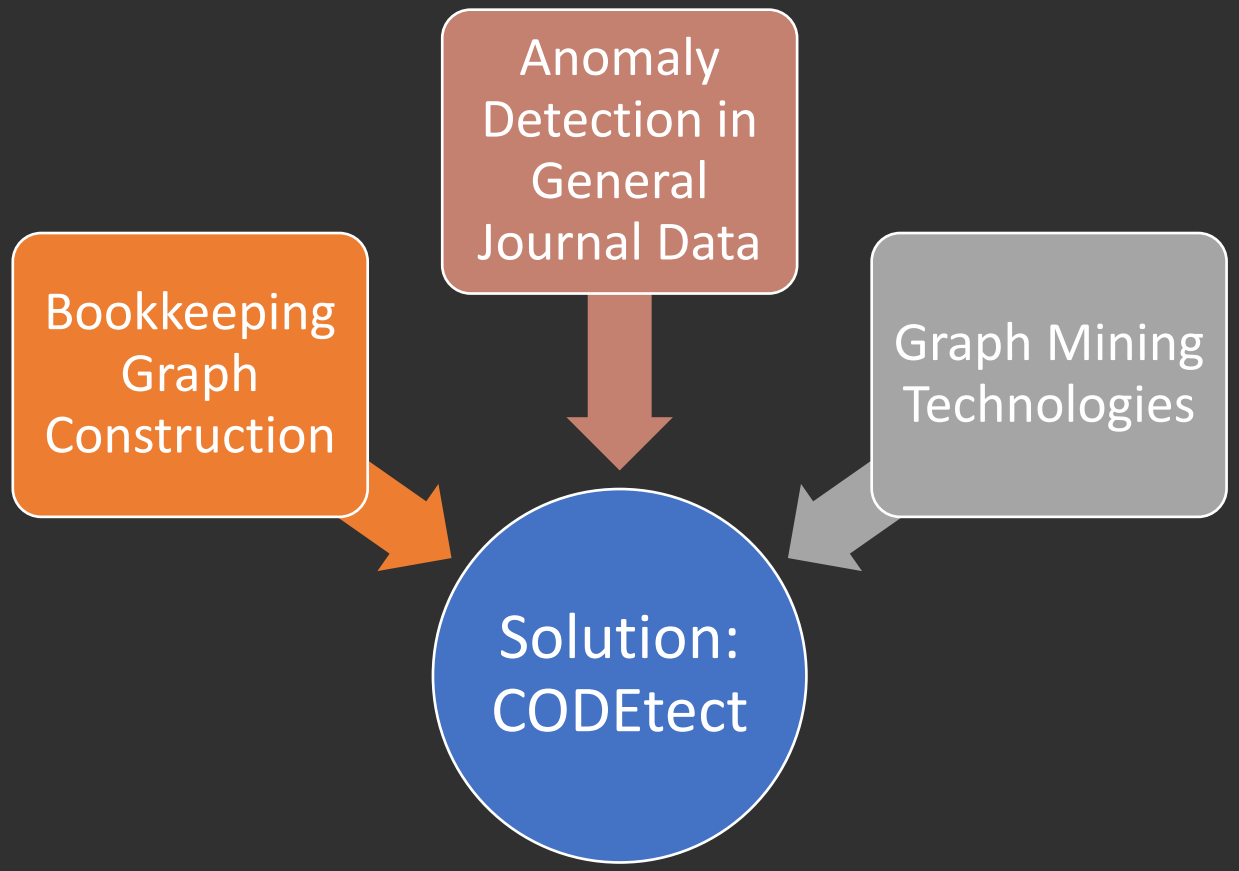
G_j

1 Mofit_3

Best Motif Table



Careful Mix Leads to Solution



Broad Lessons

- **Broader View of Research**
  - **Get closer to real world problems faced by practitioners and/or policy makers**
- **Return to Core Accounting Ideas**
  - **Double-entry bookkeeping as account graphs**
  - **Classification and aggregation**
- **Learn from Engineers**
  - **Graph theory**
  - **Graph Mining**

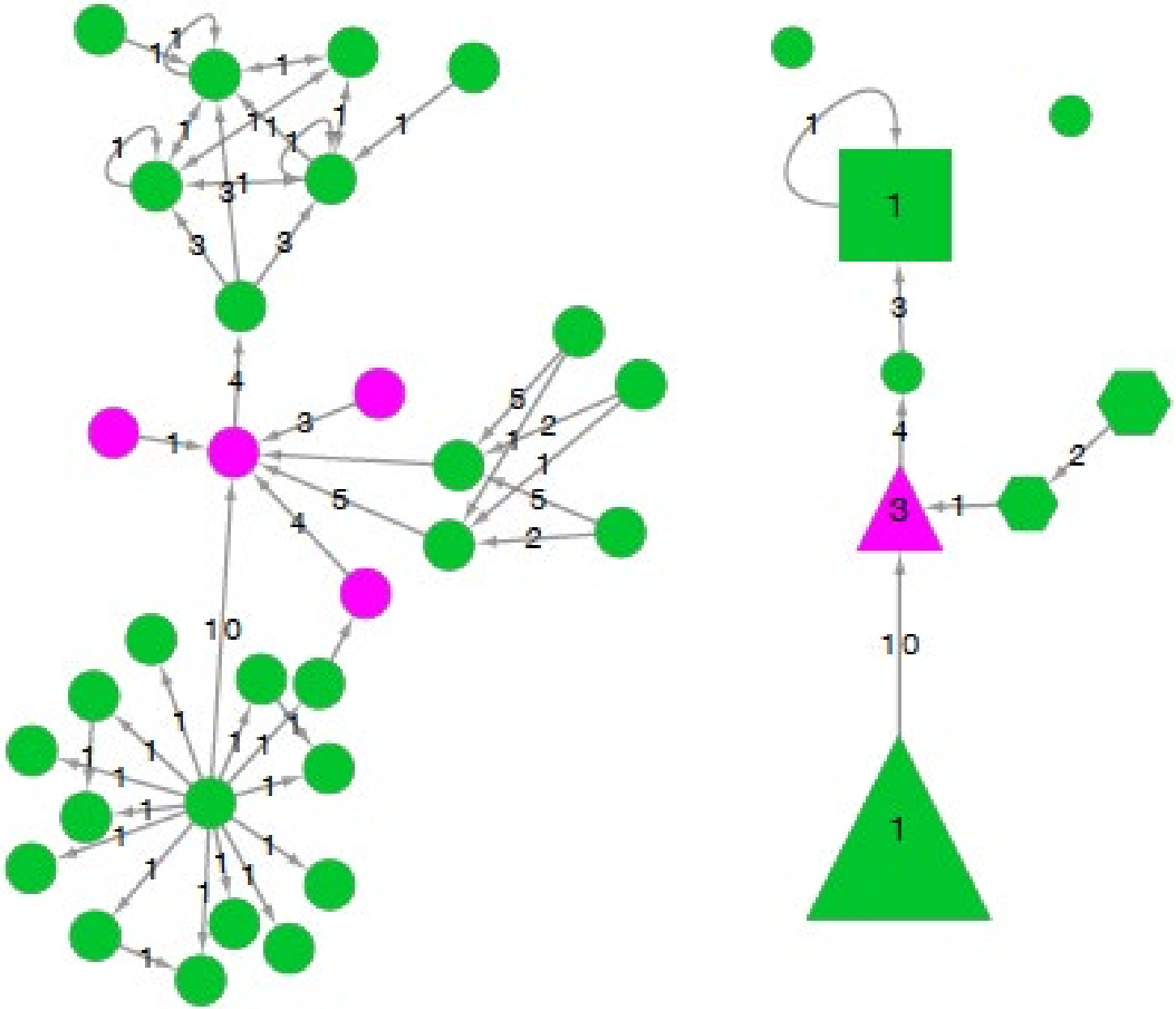
Second USE CASE:

Using Graph Summary to Quantify Aggregation in Bookkeeping Data

Principal Investigators:

Leman Akoglu and Pierre J. Liang

Carnegie Mellon University

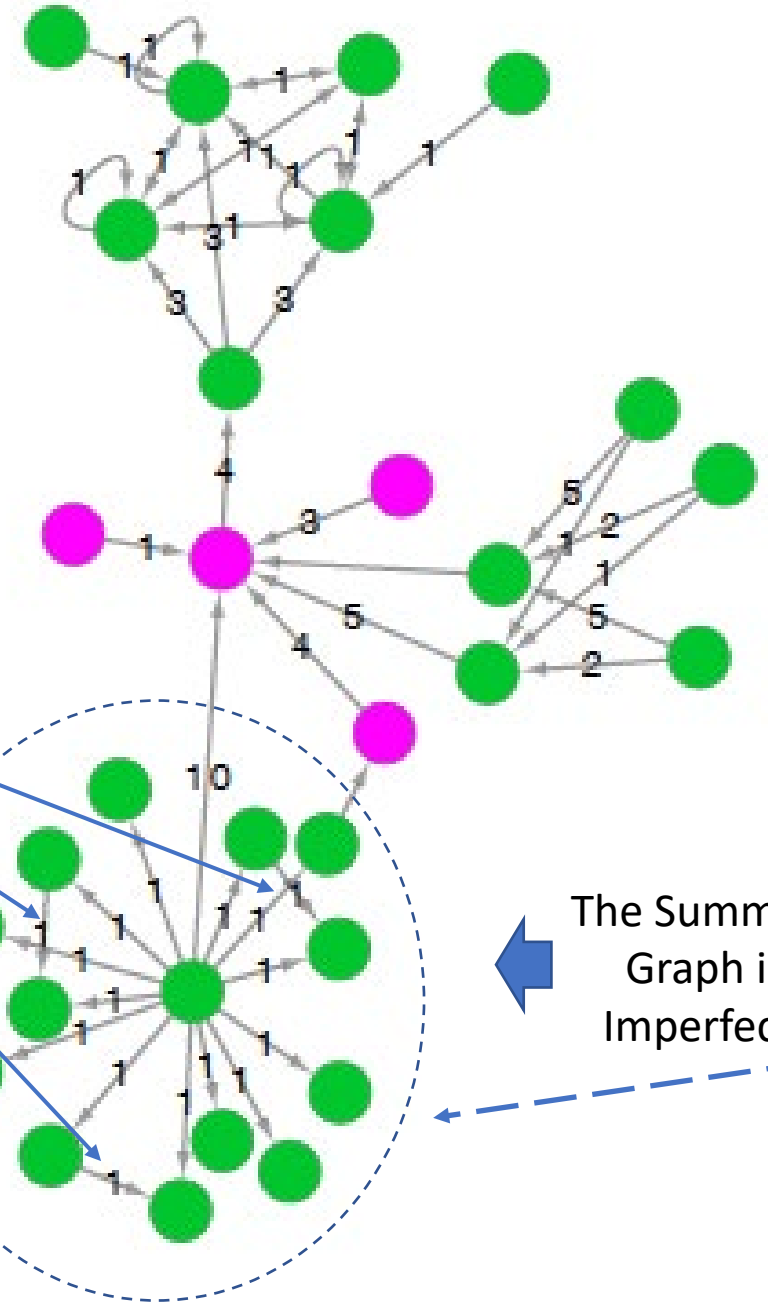

1
1
1
1
1
1
1
1
1
1
1
1
1
3
3
3
4
1
3
5
1
2
1
5
5
2
4
10
1
1
1
1
1
1
1
1
1
1
1
1

The Original Graph

G

G_S

The Summary Graph



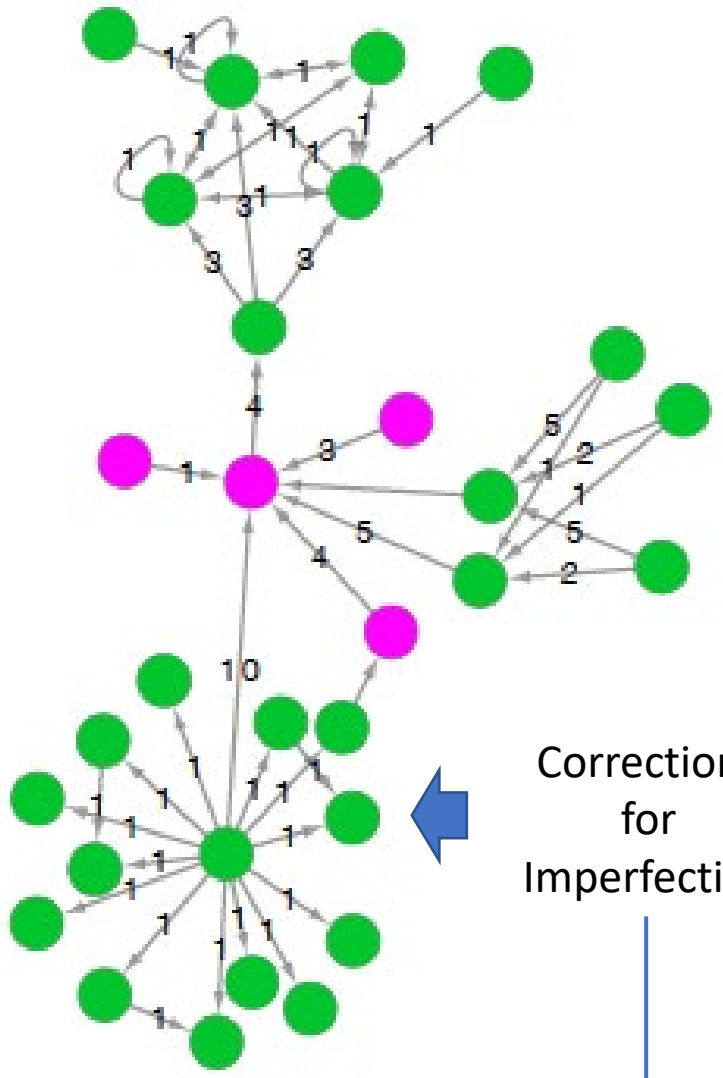
Imperfections

The Summary Graph is Imperfect!

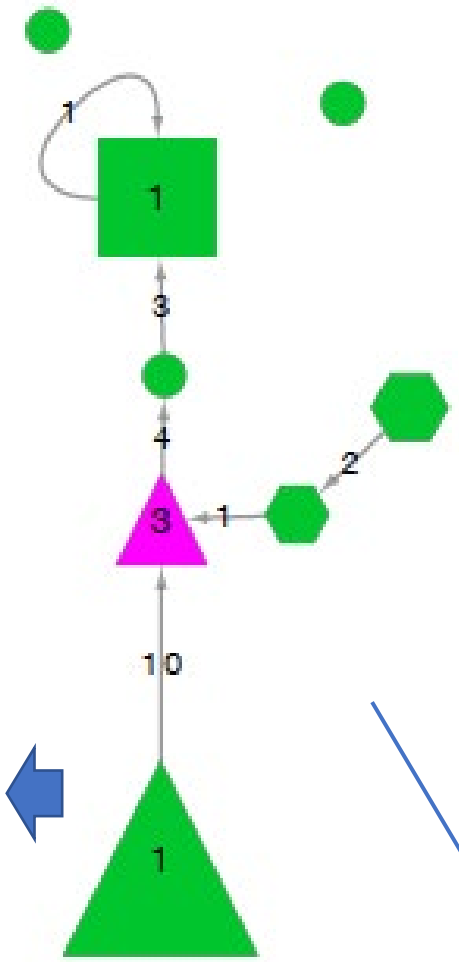
“Triangle” super-node represents a Star

The Original Graph

G



G_S



The Summary Graph

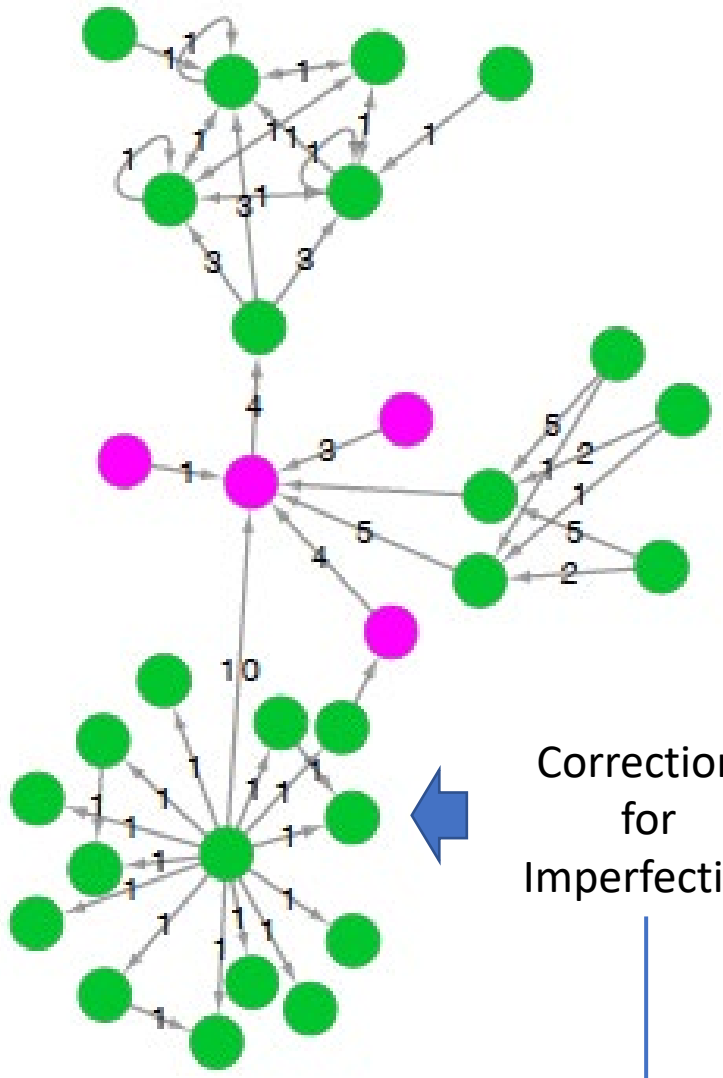
Corrections for Imperfection

Decomposition without the corrections

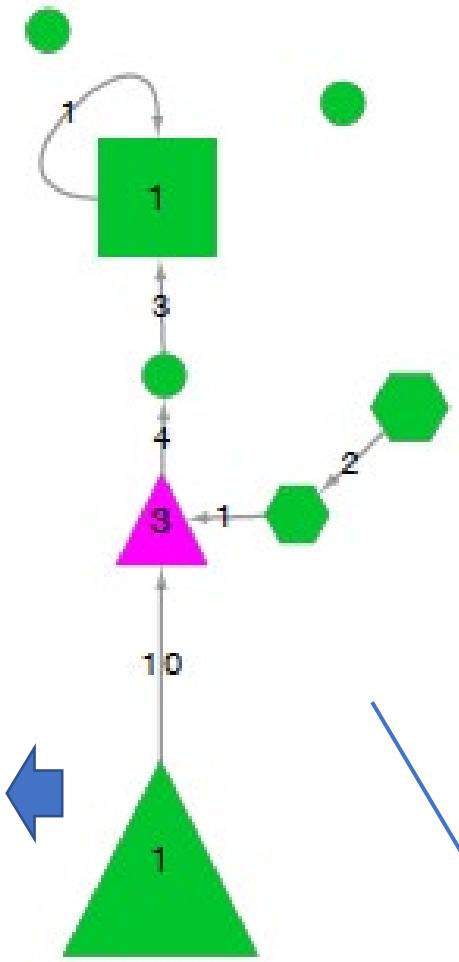
$$L(G) \geq L(G|G_S) + L(G_S)$$

The Original Graph

G



G_S



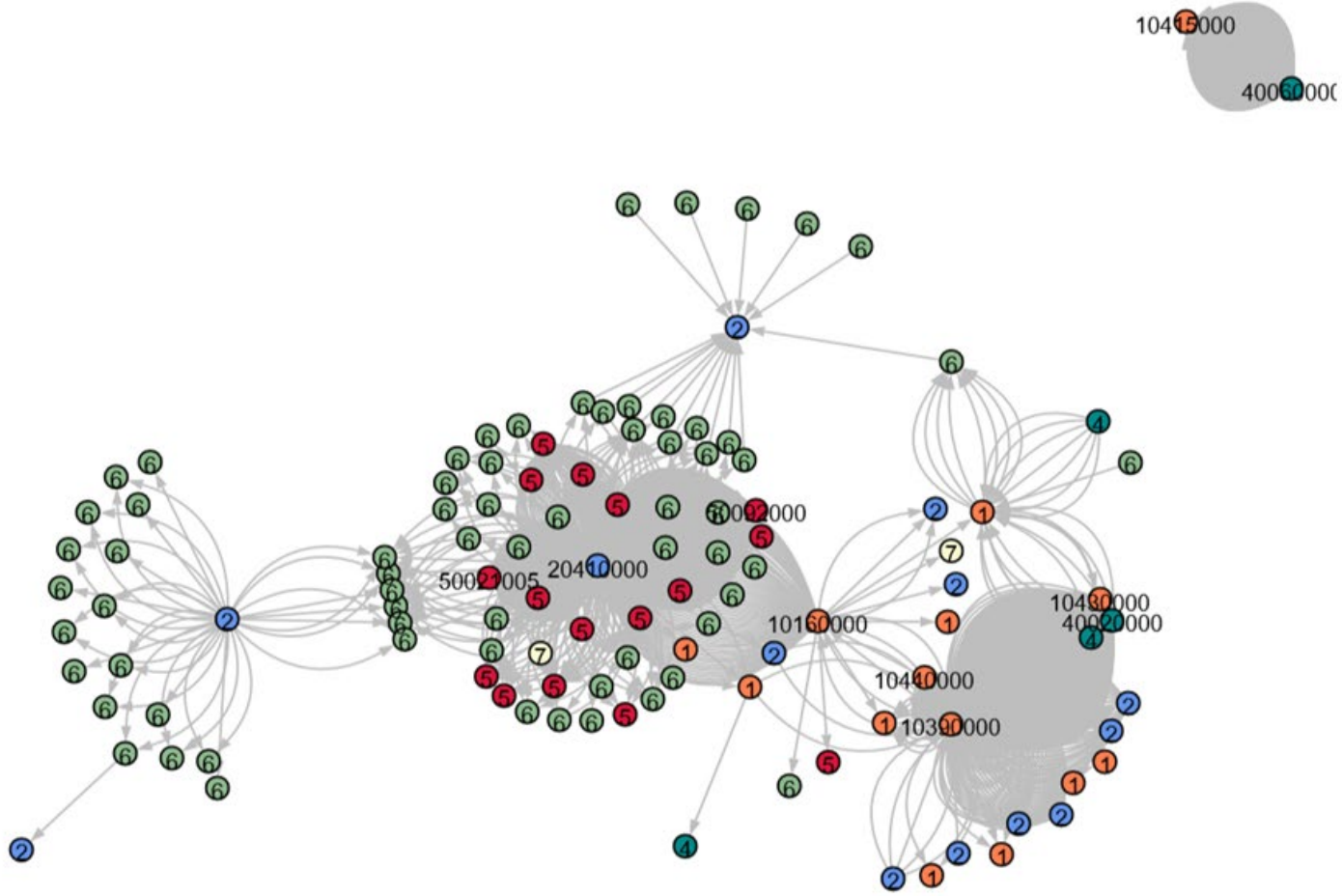
The Summary Graph

Corrections for Imperfection

Decomposition without the corrections

$$L(G) \geq L(G|G_S) + L(G_S)$$

Bookkeeping Graph for a 10-day Window



Quantifying Account Classification and Aggregation

We use TG-sum for evaluating a preexisting node labeling, the set of types pre-assigned to the nodes in a graph

Financial Statement Classification (**GAAP**) organizes the individual accounts typical GAAP categories such as inventory, prepaid expense, accounts payable, etc. For each dataset, we use specific classification given the COA given with the data.

Economic Bookkeeping Classification (**EB**) organizes the individual accounts into operating versus financing and long versus short term accounts, all together fourteen categories.

Quantifying Account Classification and Aggregation

Table 4: Evaluating account labelings in financial networks

Company	Labeling	Shuffled	Actual	norm. gain (%)
SH	EB	0.28	0.32	5.6 %
	FS	0.25	0.27	2.7 %
HW	EB	0.36	0.47	17.0 %
	FS	0.16	0.27	13.0 %
KD	EB	0.33	0.42	13.7 %
	FS	0.31	0.39	12.0 %

Many thanks!

- **Hope you learned something more about**
  - **computer science**
  - **graph theory**
  - **accounting**
- **I told you this is an unusual talk**