INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

1935-2025

90 anos de rigor e inovação
ao serviço da Sociedade

Improving the accuracy of administrative data on property transactions using a network algorithm

Alexandre Cunha
João Poças
Sofia Rodrigues
Paulo Saraiva

DRGD | SDAE

04/04/2025

Data Source – *Municipal Property Transfer Tax*

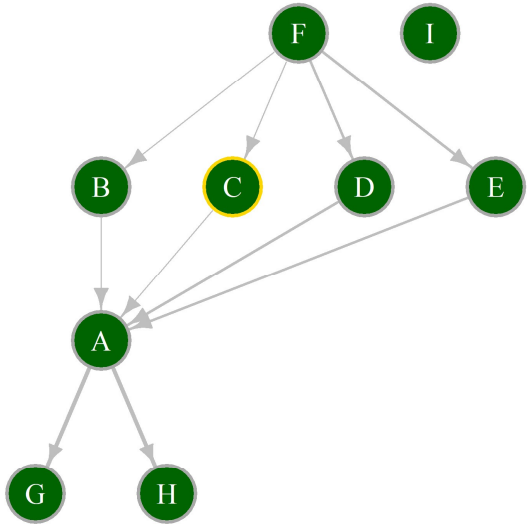
As part of the data exchange between Statistics Portugal and the Tax Authority, one of the established data flows is related to **IMT** – *Municipal Property Transfer Tax*.

IMT is a tax on the transfer of property rights in Portugal, including partial rights and equivalent transactions. The **Model 1 Declaration** is mandatory for any property transfer, detailing the buyer, co-acquirers, property, transaction details, and type of transfer (sale, exchange, or similar contracts).

Submission generates a payment reference, and IMT must be paid before the deed is signed. Errors in the declaration can be corrected upon valid request.

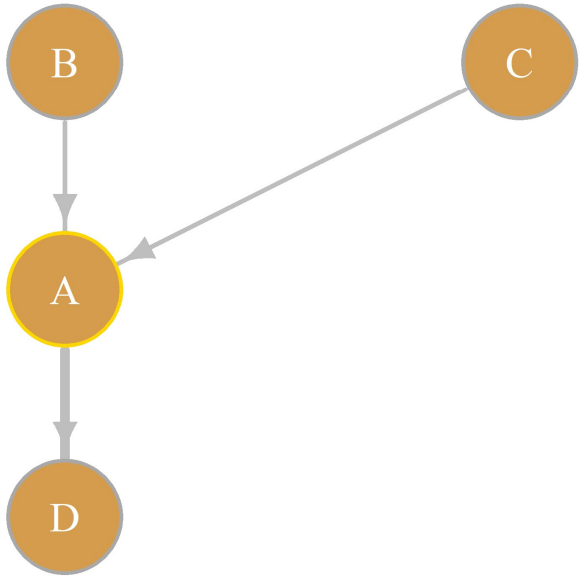
Why using network analysis?

- Network analysis often involves complex networks with many nodes and edges;
- The IMT problem is closer to linear programming than network analysis;
- However, property transactions can also be seen as unidirectional networks;
- The focus isn't on network structure but on ensuring the coherence of the unidirectional network and the consistency of edge weights.



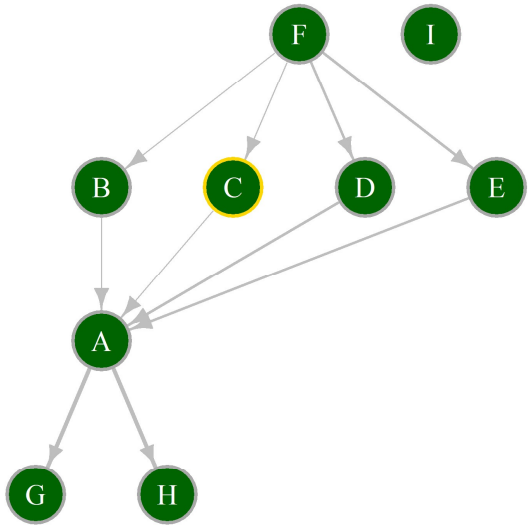
One household, one network

PROPERTY	FROM	TO	PERC	HAS_IMI	…
123	B	A	0,5	FALSE	
123	C	A	0,5	FALSE	
123	A	D	1	TRUE	
1234	B	E	1	FALSE	
56789	AA	EE	1	FALSE	
89563	FFF	GGG	0,33	TRUE	
…	…	…	…	…	…



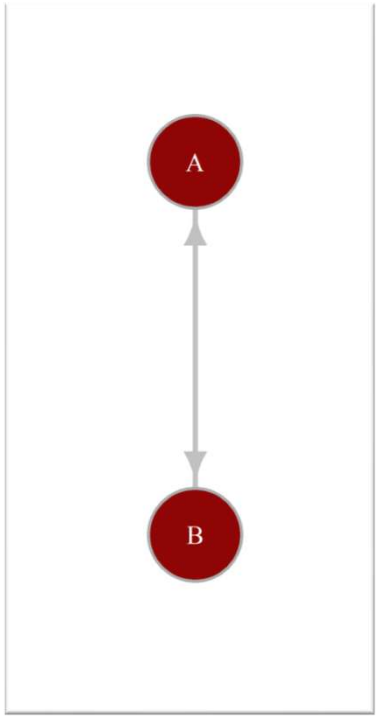
*Quo Vadis: V*alid transaction and transaction ID

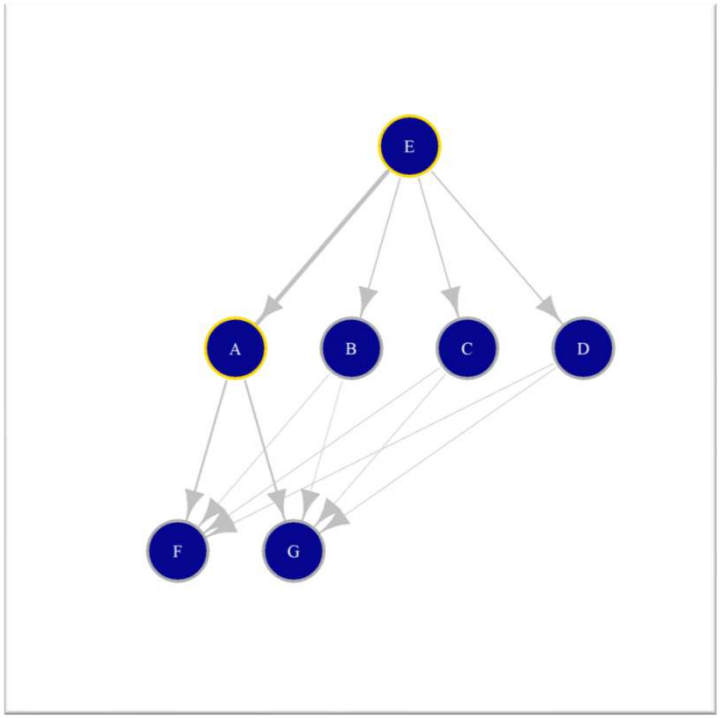
FROM	TO	PERC	FROM_IMI	TO_IMI	...	OK	ID
B	A	0,17	FALSE	FALSE		TRUE	2
C	A	0,17	TRUE	FALSE		TRUE	2
D	A	0,33	FALSE	FALSE		TRUE	2
E	A	0,33	FALSE	FALSE		TRUE	2
F	B	0,17	FALSE	FALSE		TRUE	1
F	C	0,17	FALSE	TRUE		TRUE	1
F	D	0,33	FALSE	FALSE		TRUE	1
F	E	0,33	FALSE	FALSE		TRUE	1
A	G	0,50	FALSE	FALSE		TRUE	3
A	H	0,50	FALSE	FALSE		TRUE	3
B	I	0,17	FALSE	FALSE		FALSE	
C	I	0,17	TRUE	FALSE		FALSE	
D	I	0,33	FALSE	FALSE		FALSE	
E	I	0,33	FALSE	FALSE		FALSE	

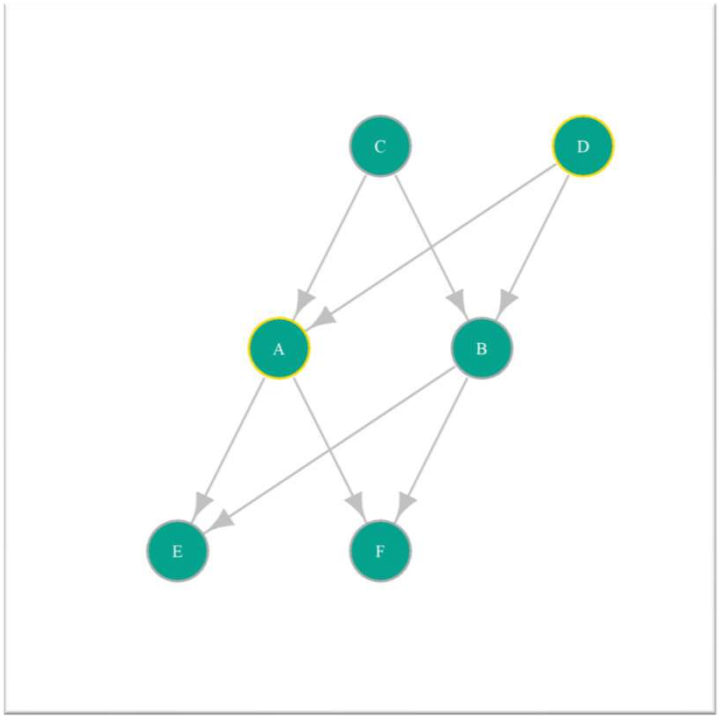


Pole Position: Picking the starting nodes

Starting nodes are those where *in* degree is zero:





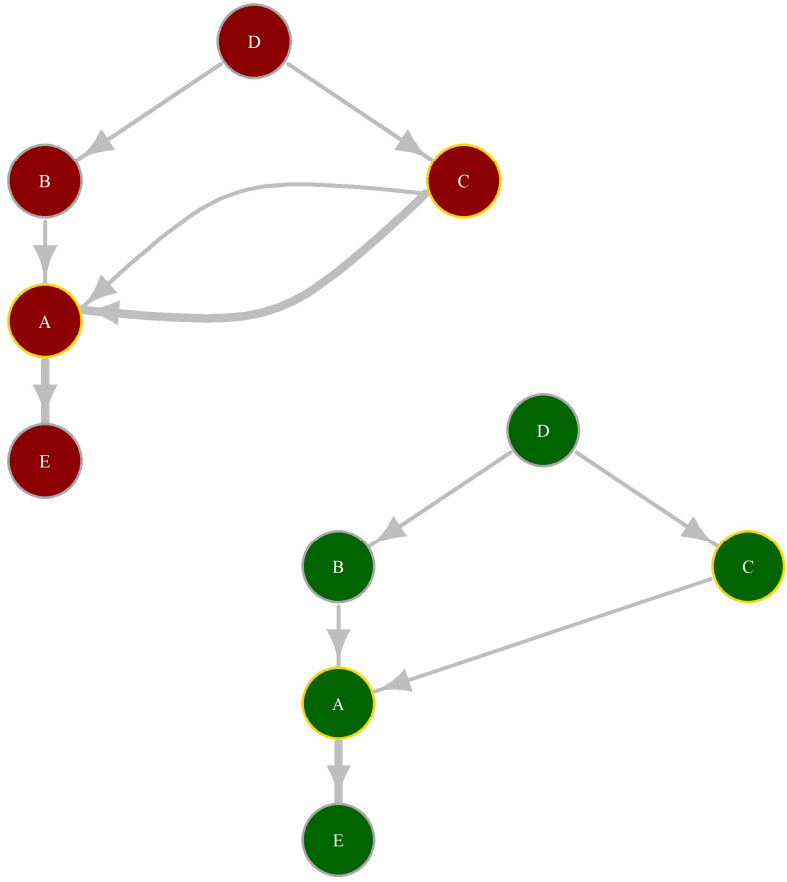


Edge in, Edge Out: Iterating through middle nodes

- Sum of ***out*** weights must be less than or equal to the sum of ***in*** weights;
- Algorithm iterates through each step of the network. Invalidated edges will not contribute to future sums;
- Edges order within each node/neighborhood defines the selection process:
  1. **Municipal Tax on Real Estate (IMI):** *does the seller have a registered property record?*
  2. **Next node *betweeness* and *degree***: *does the buyer node play an important part in the network connectivity*?
  3. **IMT order**: *we assume most recent declarations have priority over older ones.*

Edge in, Edge Out: Iterating through middle nodes

FROM	TO	PERC	FROM_IMI	TO_IMI	ORDER
A	E	1	TRUE	FALSE	1
B	A	0,5	FALSE	TRUE	2
C	A	1	TRUE	TRUE	3
C	A	0,5	TRUE	TRUE	4
D	B	0,5	FALSE	TRUE	5
D	C	0,5	FALSE	FALSE	6



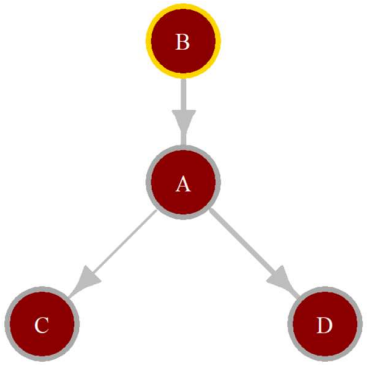
Finding Waldo: Comparison between current method and network method

We can use the current method to compare and evaluate the performance of the network:

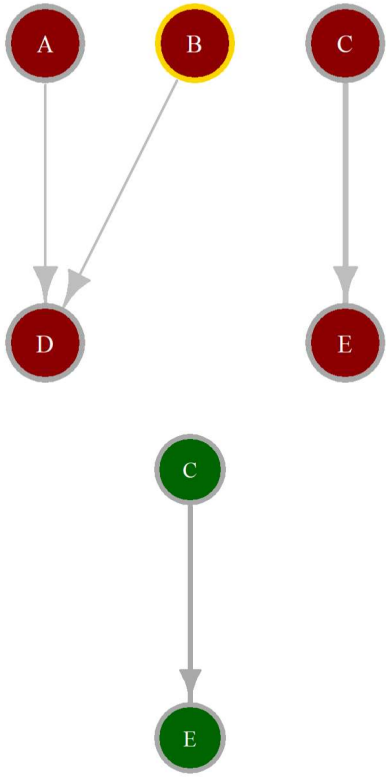
Comparison		
Result	Total	%
Pred. > Actual	139	0,37
Pred. = Actual	35 974	95,93
Pred. < Actual	1 354	3,70
Total	36 687	100,00

2nd Trimester 2024

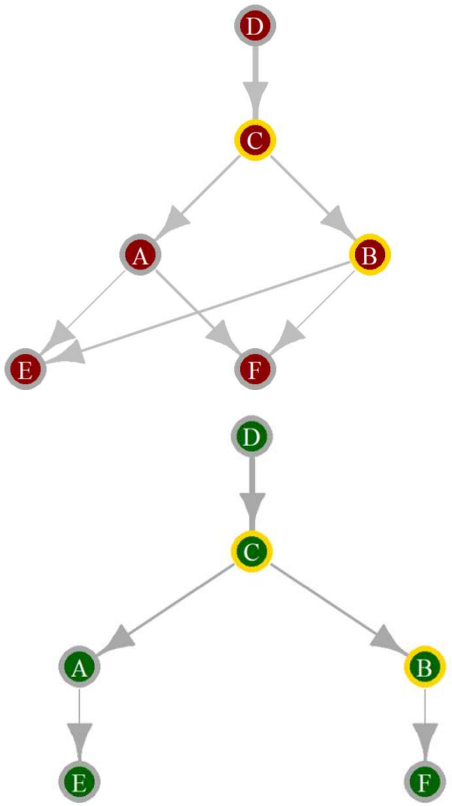
Total *network*	N *network*	Total *IPHab*	N *IPHab*	Difference
670 000	2	2 245 000	2	-1 575 000



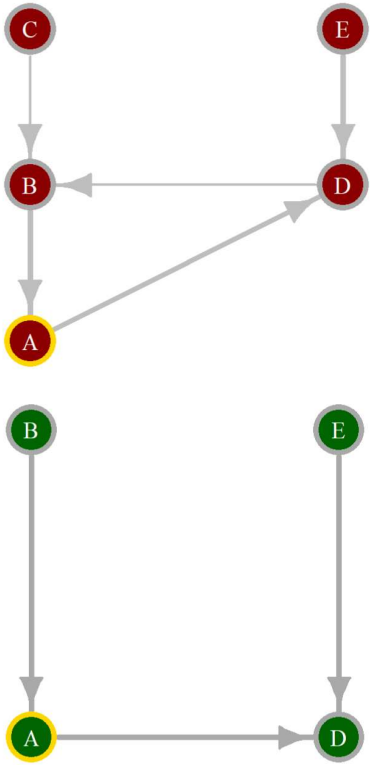
Problems yet to fix



Picking the right starting node



Potentially wrong shares



Cyclic networks

*And the winner is…*Pros and cons

PROS:

- Owners' **relations** are easly identified;
- Household **overview**

  (IMI and old IMT data);
- Powerfull **visualizations**.

CONS:

- Ignores other dataset **features**;
- **Ordering** settings can be too rigid;
- **Complex** networks can affect

  performance.

In the horizon: Next Steps

- Which is best: picking the **starting** nodes or the **ending** nodes?
- Will we need to apply **ML algorithms** when picking the best path?
- Does the algorithm have enough **quality** to be used in **production**?
- How can we measure the algorithm **robustness**, compared to the traditional algorithm?

THANK YOU

alexandre.cunha@ine.pt