

# Structure and analysis of www

Social and Technological Networks

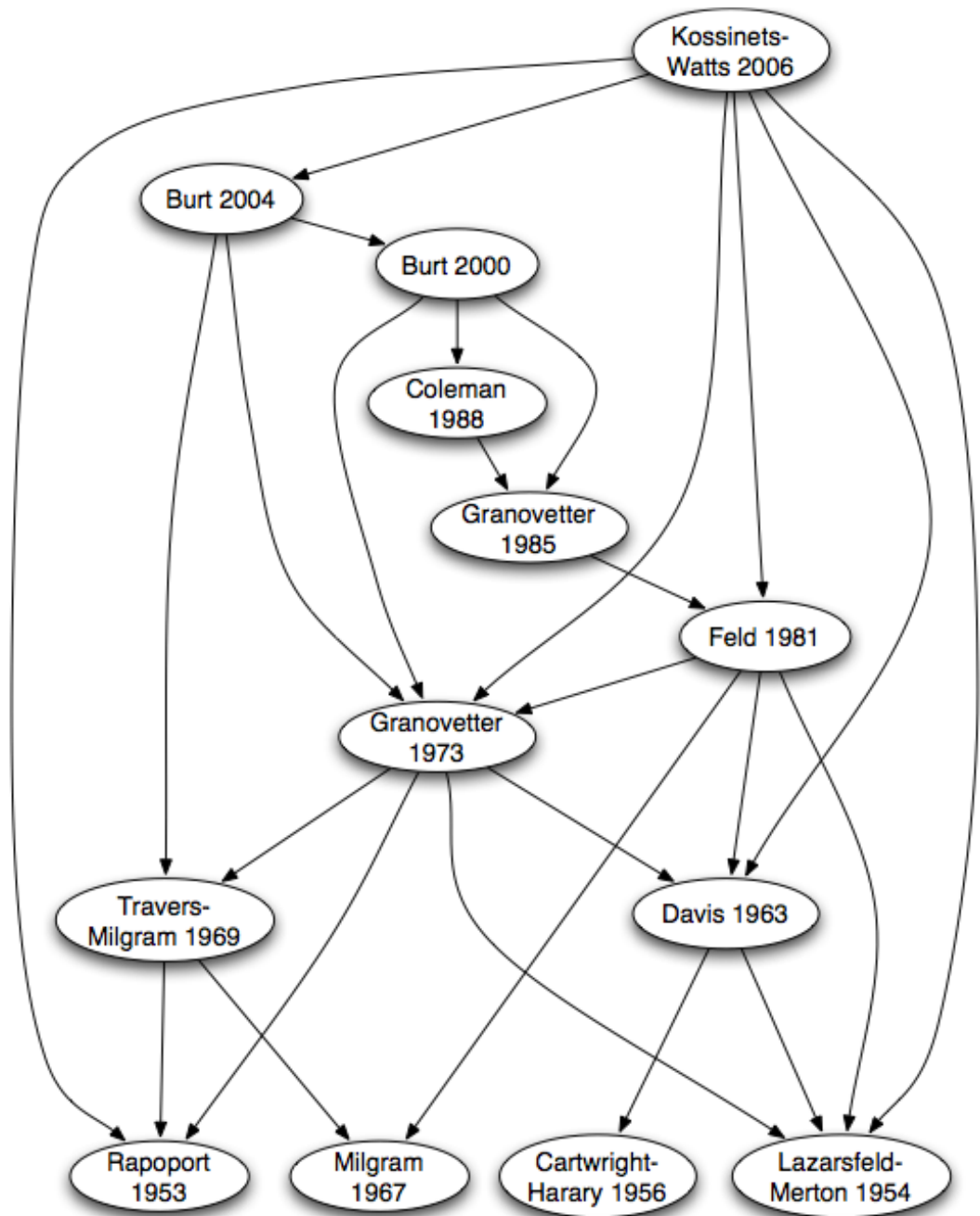
Rik Sarkar

University of Edinburgh, 2016.

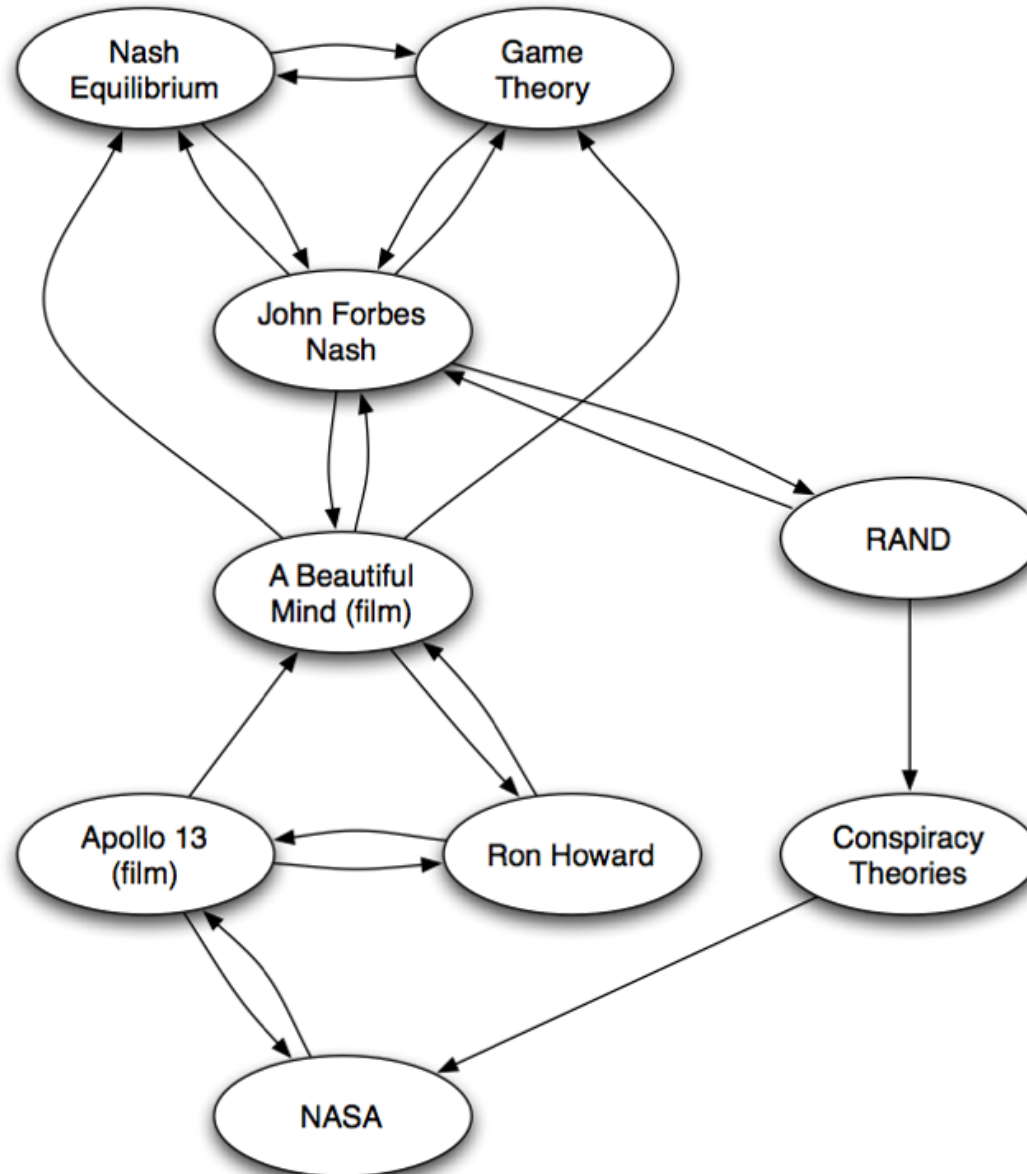
# Hyperlinks

- Give a network structure to a set of documents
  - Instead of being a simple set of documents
- Similar structure in:
  - Citations: articles, patents, legal decision,
    - Usually acyclic: citing only past documents
- Web is more dynamic — pages are updated
- not acyclic

- Example:
  - Citations



- Topics

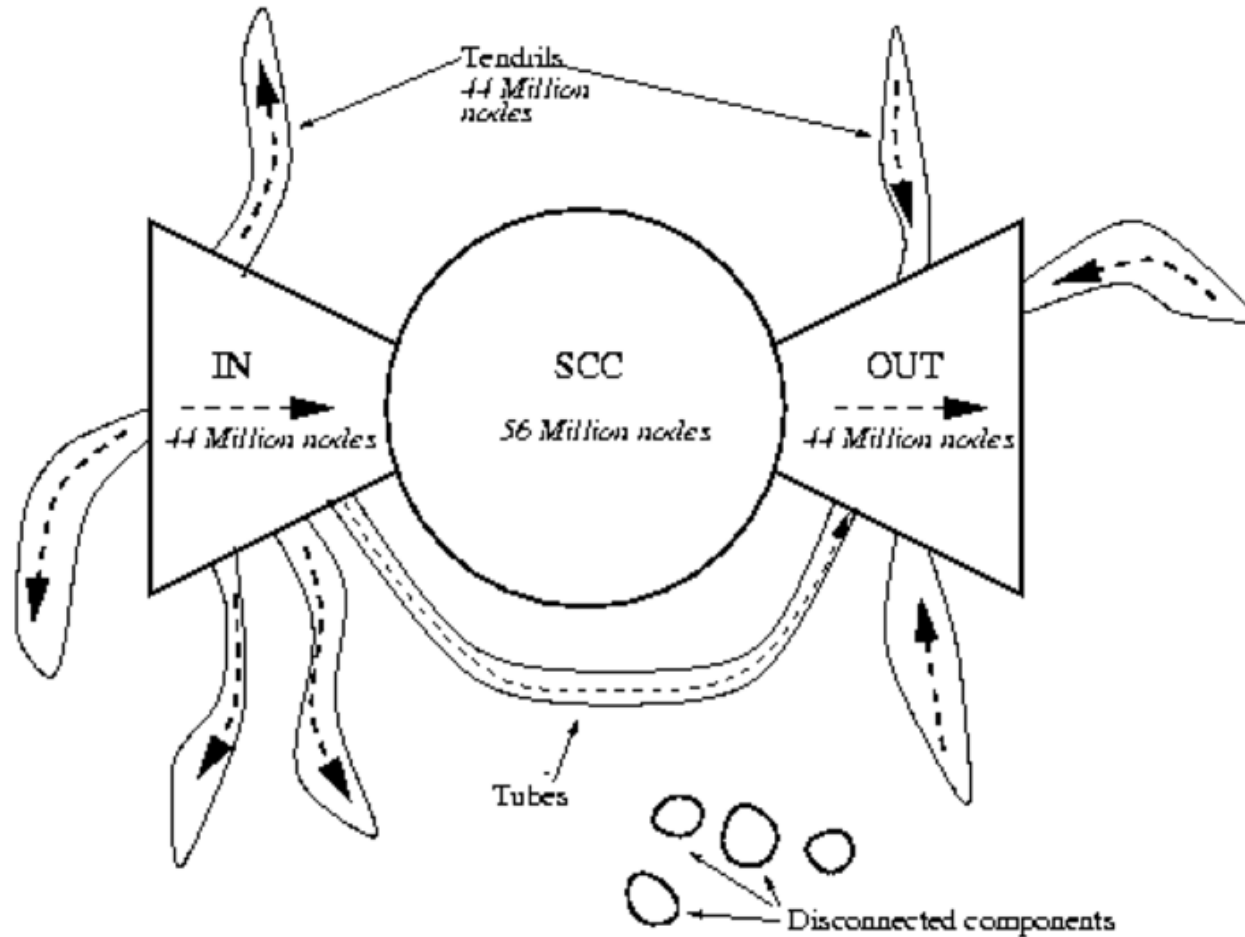


# Connected components

- In a graph:
  - A connected component is a maximal subset of nodes with a path between any pair of nodes in the subset
- In a directed graph (like the web):
  - We are interested in strongly connected components (SCC)
  - An SCC is a maximal subset of nodes, with a *directed path* between any ordered pair of nodes
    - So, there must be a path between (a, b)
    - And also between (b, a)

# Bow tie structure of the web

- Broder '99

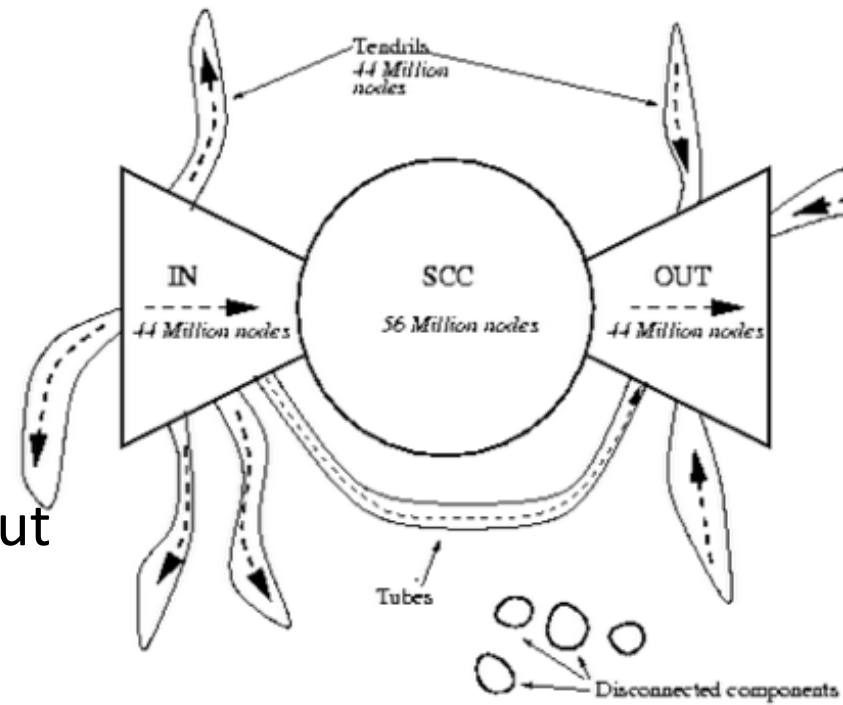


# Bow tie structure of the web

- *Single* Giant strongly connected component
- Largely due to:
  - Many topics are related to each-other (e.g. wikipedia)
  - Many search/directory sites have links to important sites, and these have links to directory/landing sites

# Bow tie structure of the web

- *Single* giant SCC
  - Hard to have 2 without links between them..
- IN nodes:
  - Flow into the GSCC
- OUT nodes:
  - Flow out of the GSCC
- Structures that do not touch GSCC
  - Tendrils: Flow into OUT and out of IN
  - Tubes: go from IN to out
  - Disconnected pieces



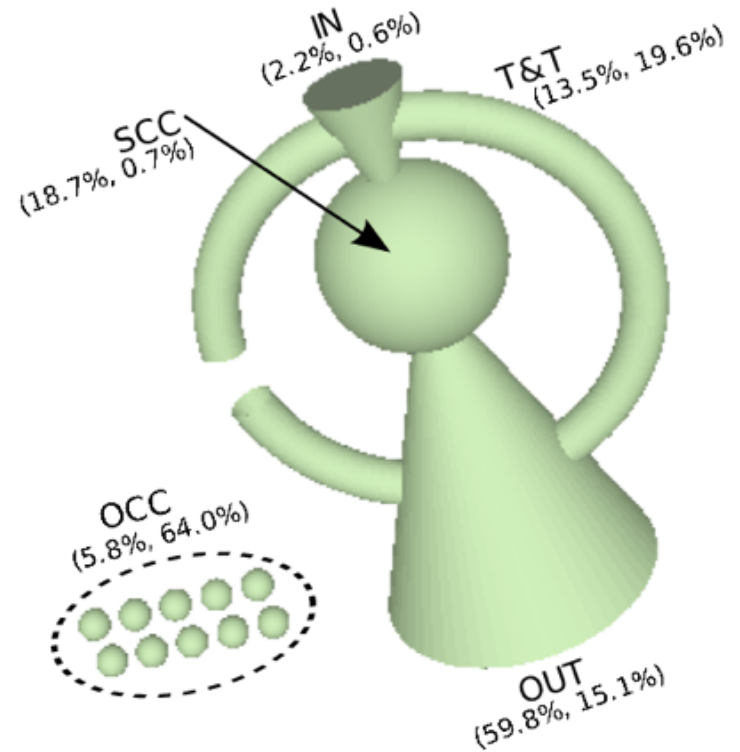


# Bow tie structure

- Similar structures in
  - Larger & recent web graphs
  - Wikipedia
  - ...

# Related: Who controls the world?

- The network of global corporate (TNC) control
- Bow tie structure
- The SCC is relatively *small*
- TNCs in SCC own most of each-other
- A group of 147 entities in SCC control About half of World's economic value
- 3/4 of the SCC are financial intermediaries



S. Vitali et al. 2011

# Searching the web

- Search for “Edinburgh” (Information retrieval)
- Find pages that match “Edinburgh”
- Decide which pages are important

About 146,000,000 results (0.59 seconds)

### The City of Edinburgh Council

[www.edinburgh.gov.uk/](http://www.edinburgh.gov.uk/)

Based in Scotland's capital city, the Council provides a range of public services to over 444000 citizens and promotes the city worldwide.

### Edinburgh & The Lothians - Scotland | VisitScotland

[www.visitscotland.com/destinations-maps/edinburgh-lothians/](http://www.visitscotland.com/destinations-maps/edinburgh-lothians/)

Welcome to Edinburgh, the inspiring capital of Scotland, where centuries of history meet a vibrant, cosmopolitan city in an unforgettable setting. Discover ...

[Things to see and do](#) - [Accommodation](#) - [Travel](#) - [About](#)

### In the news



#### Alexander Wallace named as Edinburgh hit-and-run victim

BBC News - 17 hours ago

A 57-year-old man who died in a hit-and-run in Edinburgh is named by police as a man from ...

Edinburgh street 'most polluted in Scotland'

Edinburgh Evening News - 2 hours ago

Edinburgh charity Mercy Corps help struggling families on their long walk to freedom in Lesbos

Scottish Daily Record - 5 hours ago

[More news for edinburgh](#)

### The University of Edinburgh

[www.ed.ac.uk/](http://www.ed.ac.uk/)

The University of Edinburgh, promoting excellence in teaching and research. Over 500 degree courses. One of the UK's top rated research universities. Located ...

[Postgraduate study](#) - [Undergraduate study](#) - [MyEd](#) - [Studying](#)

### Edinburgh - Wikipedia, the free encyclopedia

[en.wikipedia.org/?title=Edinburgh](http://en.wikipedia.org/?title=Edinburgh)

Edinburgh has been recognised as the capital of Scotland since at least the 16th .... In Edinburgh, the Town Council, keen to emulate London by initiating city ...

[Edinburgh Castle](#) - [List of towns and cities in ...](#) - [Lothian](#) - [University of Edinburgh](#)

### Things To Do and See In Edinburgh, This is Edinburgh

[thisisedinburgh.com/](http://thisisedinburgh.com/)

This is Edinburgh, A Hub of All the Best Things to See and Do in Edinburgh. Explore our city.

### Images for edinburgh

[Report images](#)



[More images for edinburgh](#)

### Edinburgh fringe festival 2015: what to see and ...

[www.theguardian.com/.../edinburgh-festival-2015-what-to-...](http://www.theguardian.com/.../edinburgh-festival-2015-what-to-...) - The Guardian

So the 2015 Edinburgh fringe programme is finally out. The days when I waded like a terrier for ... Edinburgh fringe 2015 lineup: grab your chance to see theatre's future. Lyn Gardner.



## Edinburgh

Capital of Scotland

Hilly Edinburgh, Scotland's capital, has a medieval Old Town and an elegant Georgian New Town, with gardens and neoclassical buildings. It's home to Arthur's Seat, an extinct volcano in Holyrood Park with sweeping views from its peak. Looming over the city is hilltop Edinburgh Castle, home to Scotland's crown jewels and the Stone of Destiny, traditionally used in the coronation of Scottish rulers.

**Weather:** 9°C, Wind W at 3 mph (5 km/h), 73% Humidity

**Population:** 492,680 (2014)

**Local time:** Tuesday 10:39

### Upcoming events

Midas Fall	Sat 17 Oct
Rudimental Edinburgh Corn Exchange	Tue 13 Oct
Johnny Marr The Liquid Room	Wed 14 Oct

### Points of interest

[View 10+ more](#)



Edinburgh Castle



Holyrood Palace



Royal Botanic Garden E...



HMY Britannia



Arthur's Seat

[More about Edinburgh](#)

[Feedback](#)

# Searching the web

- How do you decide:
  - University of Edinburgh is more important than
  - Edinburgh dry-cleaners
- Analyze the web graph to see which node is more important

# The basic idea

- In-links constitute a vote for importance
  - If somebody is linking to a web page, that means they see something of value in it
  - If many people are linking to it, then likely the page is valuable to many other people as well

# Enhanced idea

- Not all links imply equal importance
- Links from *Important* pages are more valuable than links from unimportant pages
- Thus, we have an iterative idea:
  1. Decide importance of pages
  2. Update importance of their neighbors suitably
  3. Repeat

# The HITS algorithm

- Not all pages are similar
- Some are important for the information they contain (Authorities) (e.g. course pages)
- Some are important for the links they contain (Hubs) (e.g. list of courses)
  - They guide you to the right authorities
- Let's rank them separately, but depending on each other
  - A hub linking to good authorities is likely good
  - An authority linked by good hubs is likely good



# Hubs and authorities

- For each page  $p$ , estimate its score both as:
  - A hub:  $hub(p)$
  - An authority:  $auth(p)$
- Repeatedly in each round

# Update rules

- Start with all hub and auth = 1
- Apply Authority update to all nodes:
  - $\text{auth}(p) = \text{sum of all hub}(q) \text{ where } q \rightarrow p \text{ is a link}$
- Apply Hub update to all nodes:
  - $\text{hub}(p) = \text{sum of all auth}(r) \text{ where } p \rightarrow r \text{ is a link}$
- Repeat for k rounds

# Normalize

- We need only relative values.
- Divide each  $auth(p)$  by sum of all  $auth$  scores
- Divide each  $hub(p)$  by sum of all  $hub$  scores

# Pagerank

- Idea: Not all pages have good classification as hubs/authorities
- Sometimes authorities link directly to each other
  - Eg. wikipedia pages

# Pagerank: basic algorithm

- Overall “value” in the system is conserved = 1
- Assign “value”  $1/n$  to each node
- In each round
  - Each node divides equal portion of its pagerank value to its out-going links
  - Updates its own value to be sum of values it receives

What are the difficulties of pagerank?

# What are the difficulties of pagerank?

- Acyclic graph:
- Some nodes can get all the values
  - Lakes/seas at the local minima
- Some nodes can end without any value
  - Mountains or peaks (maxima)

# Scaled pagerank

- In every round:
  - Divide  $s$  fraction of your pagerank equally among neighbors
  - Divide  $(1-s)$  fraction equally among all nodes in the network



# The random-walk interpretation

- Users start at random web pages
- Then click links on them randomly
- Sometimes (with  $P_r = 1-s$ ) they decide to leave the page and jump to a random page in the web

# Other improvements

- Use textual information
- Use usage data: which links people click
- Use other contextual data
  - Location, personal history etc...
- Adjustment to SEO
- Adaptation to the fast changing web...

# Properties

- HITS converges
- Pagerank Converges
- Pagerank is equivalent to random walk

# References

- Please read:
- Chapter 13 & 14 in Kleinberg & Easley
- Including advanced material in ch 14.
- We will cover that in class

# About projects

- Designing & implementing new algorithms is good
- Experimenting with existing ones is also fine
- Use existing libraries if convenient
- You are free to do anything. End goal is to produce some interesting results
  - Think about special cases:
    - What happens for small world graphs? Power law? Road networks?
  - Compare algorithms
  - There is no “ideal” answer
- Discuss with us and others. (we don't know the answers, but can give some thoughts)
- Office hours (Rik): Wednesdays 3:30.

# Adjacency Matrix

- Work this out on your own and see if it makes sense:
- $M(i,j) = 1$  iff there is an edge  $i \rightarrow j$
- $M(i,j) = 0$  otherwise
- Now suppose  $\mathbf{a}$  is the vector of authority values
- Then the hub update rule is equivalent to:
  - $\mathbf{h} := \mathbf{M}\mathbf{a}$