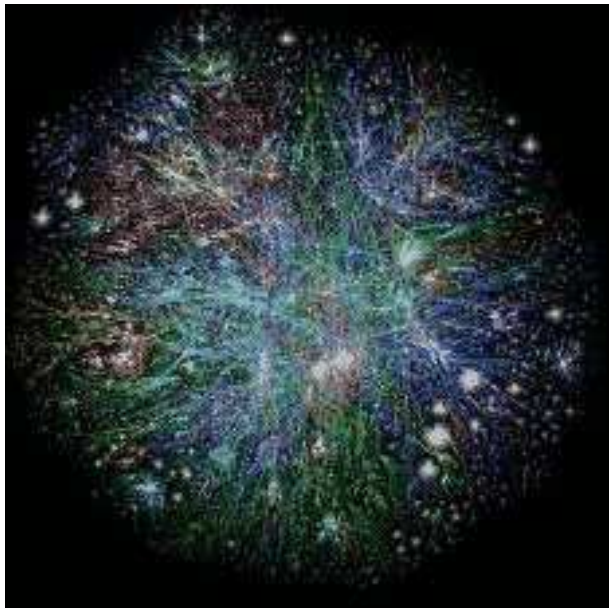
The Mathematics of the WEB

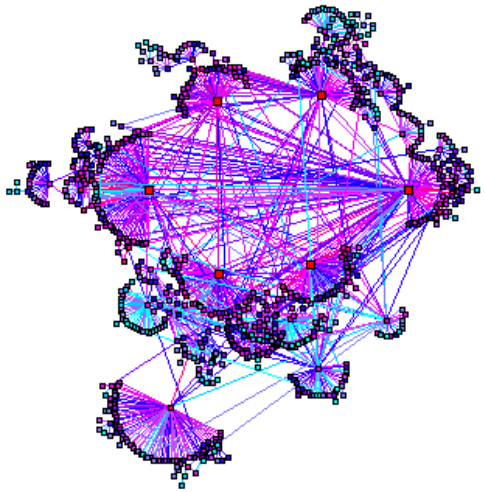
Edmundo de Souza e Silva[1]
Daniel Sadoc Menasche[2]

Federal University of Rio de Janeiro
[1]Systems Engineering and Computer Science Department, COPPE
[2]Computer Science Department, Math Institute

2011

Internet





- The Internet and the Web: full of intersting real problems
- Needs mathematical tools to understand/solve these

Outline

Some Problems

- when you type a query how a search engine (like Google) chooses what to show first?
- Since the web is dynamic, how to efficiently craw the web and to decide which pages to update?
- Which metrics are appropriate to capture important network characteristics and how to calculate them? (how to sample?)
- How to determine the most *influential* node in a network?
- How to discover if there is a path from "s"to "d"?
- How to efficiently search for information and retrieve it (e.g. a movie)?

Some Problems

- when you type a query how a search engine (like Google) chooses what to show first?
- Since the web is dynamic, how to efficiently craw the web and to decide which pages to update?
- Which metrics are appropriate to capture important network characteristics and how to calculate them? (how to sample?)
- How to determine the most *influential* node in a network?
- How to discover if there is a path from "s"to "d"?
- How to efficiently search for information and retrieve it (e.g. a movie)?

Some Problems

- when you type a query how a search engine (like Google) chooses what to show first?
- Since the web is dynamic, how to efficiently craw the web and to decide which pages to update?
- Which metrics are appropriate to capture important network characteristics and how to calculate them? (how to sample?)
- How to determine the most *influential* node in a network?
- How to discover if there is a path from "s"to "d"?
- How to efficiently search for information and retrieve it (e.g. a movie)?

Some Problems

- when you type a query how a search engine (like Google) chooses what to show first?
- Since the web is dynamic, how to efficiently craw the web and to decide which pages to update?
- Which metrics are appropriate to capture important network characteristics and how to calculate them? (how to sample?)
- How to determine the most *influential* node in a network?
- How to discover if there is a path from "s"to "d"?
- How to efficiently search for information and retrieve it (e.g. a movie)?

Some Problems

- when you type a query how a search engine (like Google) chooses what to show first?
- Since the web is dynamic, how to efficiently craw the web and to decide which pages to update?
- Which metrics are appropriate to capture important network characteristics and how to calculate them? (how to sample?)
- How to determine the most *influential* node in a network?
- How to discover if there is a path from "s"to "d"?
- How to efficiently search for information and retrieve it (e.g. a movie)?

Some Problems

- when you type a query how a search engine (like Google) chooses what to show first?
- Since the web is dynamic, how to efficiently craw the web and to decide which pages to update?
- Which metrics are appropriate to capture important network characteristics and how to calculate them? (how to sample?)
- How to determine the most *influential* node in a network?
- How to discover if there is a path from "s"to "d"?
- How to efficiently search for information and retrieve it (e.g. a movie)?

Objective of the lecture

- To present a few recent practical problems of the WEB
- To show the mathematical foundations behind the solutions for the problems
- To motivate the students on the subject area
- The lecture is NOT:
  - To present solutions to the problems
  - To develop the mathematical tools

Objective of the lecture

- To present a few recent practical problems of the WEB
- To show the mathematical foundations behind the solutions for the problems
- To motivate the students on the subject area
- The lecture is NOT:
  - To present solutions to the problems
  - To develop the mathematical tools

Objective of the lecture

- To present a few recent practical problems of the WEB
- To show the mathematical foundations behind the solutions for the problems
- To motivate the students on the subject area
- The lecture is NOT:
  - To present solutions to the problems
  - To develop the mathematical tools

Objective of the lecture

- To present a few recent practical problems of the WEB
- To show the mathematical foundations behind the solutions for the problems
- To motivate the students on the subject area
- The lecture is NOT:
  - To present solutions to the problems
  - To develop the mathematical tools

Objective of the lecture

- To present a few recent practical problems of the WEB
- To show the mathematical foundations behind the solutions for the problems
- To motivate the students on the subject area
- The lecture is NOT:
  - To present solutions to the problems
  - To develop the mathematical tools

Objective

- Touch on topics such as:
  - Page Ranking
  - Markov Chains
  - Random Walk on Graphs
  - Measurements: Centrality, Clustering, etc
  - Reacommendation Systems
  - P2P scalability
- Cover Application examples, not theory!

Objective

- Touch on topics such as:
  - Page Ranking
  - Markov Chains
  - Random Walk on Graphs
  - Measurements: Centrality, Clustering, etc
  - Reacommendation Systems
  - P2P scalability
- Cover Application examples, not theory!

Objective

- Touch on topics such as:
  - Page Ranking
  - Markov Chains
  - Random Walk on Graphs
  - Measurements: Centrality, Clustering, etc
  - Reacommendation Systems
  - P2P scalability
- Cover Application examples, not theory!

Ranking Pages

- When you type a query, what happens?
- Search engines return a list of URL's based on a set of key words describing the interest of the client.
- Search engine decides which page to show first, second, etc.
- Problem: how to rank web pages.
- Assign a ranking to pages which is some measure of the quality, or importance of that page . . . $\rightarrow$ the Google search engine

Ranking Pages

- When you type a query, what happens?
- Search engines return a list of URL's based on a set of key words describing the interest of the client.
- Search engine decides which page to show first, second, etc.
- Problem: how to rank web pages.
- Assign a ranking to pages which is some measure of the quality, or importance of that page . . . $\rightarrow$ the Google search engine

Ranking Pages

- When you type a query, what happens?
- Search engines return a list of URL's based on a set of key words describing the interest of the client.
- Search engine decides which page to show first, second, etc.
- Problem: how to rank web pages.
- Assign a ranking to pages which is some measure of the quality, or importance of that page . . . $\rightarrow$ the Google search engine

Ranking Pages

- When you type a query, what happens?
- Search engines return a list of URL's based on a set of key words describing the interest of the client.
- Search engine decides which page to show first, second, etc.
- Problem: how to rank web pages.
- Assign a ranking to pages which is some measure of the quality, or importance of that page . . . $\rightarrow$ the Google search engine

Ranking Pages

- When you type a query, what happens?
- Search engines return a list of URL's based on a set of key words describing the interest of the client.
- Search engine decides which page to show first, second, etc.
- Problem: how to rank web pages.
- Assign a ranking to pages which is some measure of the quality, or importance of that page ... $\rightarrow$ the Google search engine

Ranking Pages
A Graph

- A web page has both forward and backward links. $\rightarrow$ correspondence to a directed graph
- Graph?
  - represents relationships between pairs of objects in a set.
  - Ex: objects are pages, relationship: link between pages
  - Ex: objects are network routers, relationship: data channel connecting routers
  - Many other examples

Ranking Pages
A Graph

- A web page has both forward and backward links. $\rightarrow$ correspondence to a directed graph
- Graph?
  - represents relationships between pairs of objects in a set.
  - Ex: objects are pages, relationship: link between pages
  - Ex: objects are network routers, relationship: data channel connecting routers
  - Many other examples

Ranking Pages
A Graph

- A web page has both forward and backward links. $\rightarrow$ correspondence to a directed graph
- Graph?
  - represents relationships between pairs of objects in a set.
  - Ex: objects are pages, relationship: link between pages
  - Ex: objects are network routers, relationship: data channel connecting routers
  - Many other examples

Ranking Pages
A Graph

- A web page has both forward and backward links. $\rightarrow$ correspondence to a directed graph
- Graph?
  - represents relationships between pairs of objects in a set.
  - Ex: objects are pages, relationship: link between pages
  - Ex: objects are network routers, relationship: data channel connecting routers
  - Many other examples

Ranking Pages
A Graph

- A web page has both forward and backward links. $\rightarrow$ correspondence to a directed graph
- Graph?
  - represents relationships between pairs of objects in a set.
  - Ex: objects are pages, relationship: link between pages
  - Ex: objects are network routers, relationship: data channel connecting routers
  - Many other examples

Ranking Pages
A Graph

- A web page has both forward and backward links. $\rightarrow$ correspondence to a directed graph
- Graph?
  - represents relationships between pairs of objects in a set.
  - Ex: objects are pages, relationship: link between pages
  - Ex: objects are network routers, relationship: data channel connecting routers
  - Many other examples

Ranking Pages

- Naive approach: assign the rank of a page as the number of links to that page from other pages.
- Problem???
- For instance, one can artificially increase the rank of a page by creating a large number of web pages that link to that page.

Ranking Pages

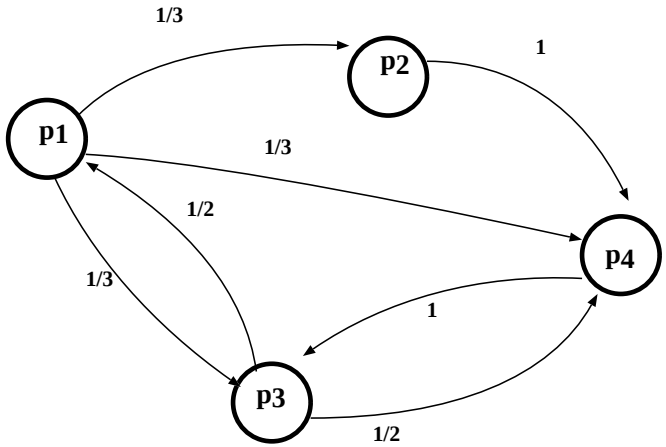
- Naive approach: assign the rank of a page as the number of links to that page from other pages.
- Problem???
- For instance, one can artificially increase the rank of a page by creating a large number of web pages that link to that page.

Ranking Pages

- Naive approach: assign the rank of a page as the number of links to that page from other pages.
- Problem???
- For instance, one can artificially increase the rank of a page by creating a large number of web pages that link to that page.

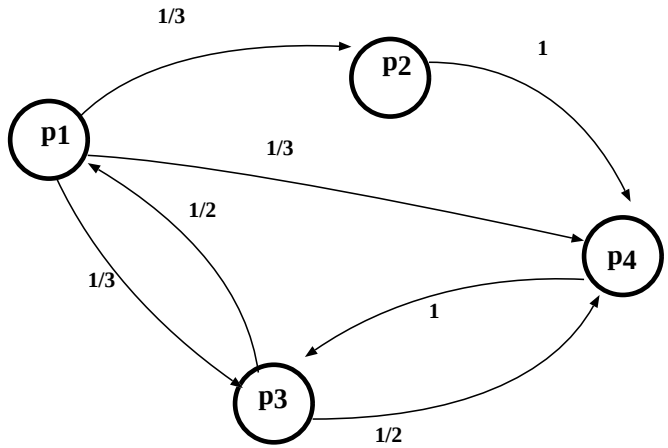
Random Surfer model

- Idea: the rank of a page should derive not just from the number of links to that page but somehow weighted with the ranks of those pages that point to the page.
- Original Google approach: the contribution a page makes to the ranking of the pages it links to should be split evenly among the pages to which it links.

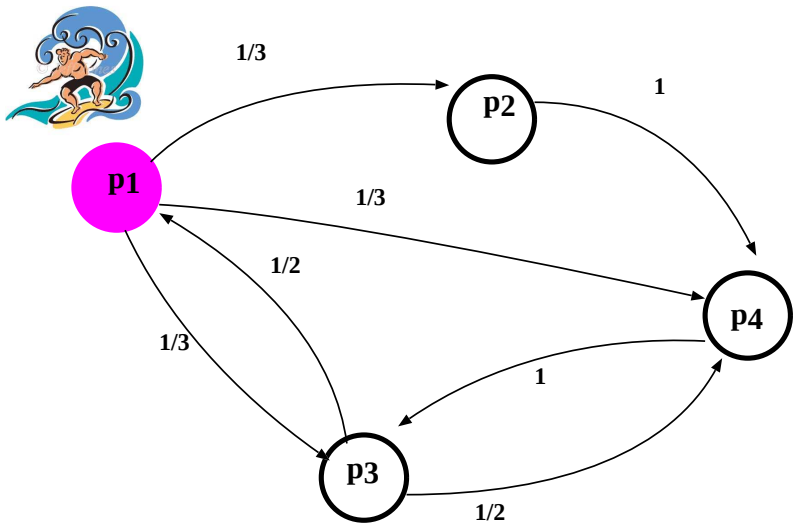


Random Surfer model

- Idea: the rank of a page should derive not just from the number of links to that page but somehow weighted with the ranks of those pages that point to the page.
- Original Google approach: the contribution a page makes to the ranking of the pages it links to should be split evenly among the pages to which it links.



Random Surfer model

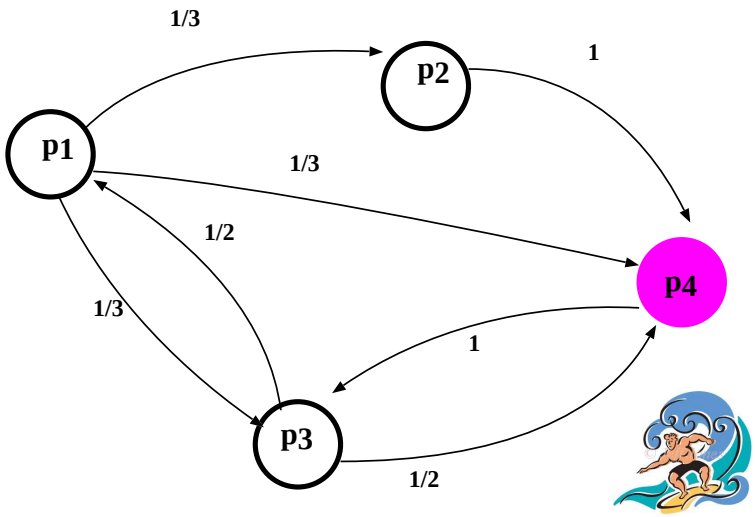


	# of Visits	Fraction of Visits	Probability of Visit
1	1	1	1
2	0	0	0
3	0	0	0
4	0	0	0

Random Walker | Power Method

$$\begin{bmatrix} 0 & 0.33 & 0.33 & 0.33 \\ 0 & 0 & 0 & 1 \\ 0.5 & 0 & 0 & 0.5 \\ 0 & 0 & 1 & 0 \end{bmatrix}^{0}$$

Random Surfer model

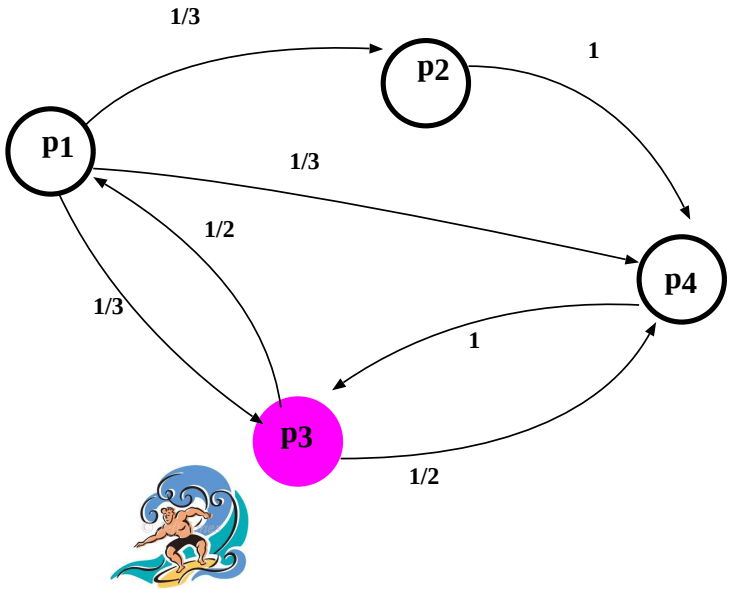


	# of Visits	Fraction of Visits	Probability of Visit
1	1	0.5	0
2	0	0	0.33
3	0	0	0.33
4	1	0.5	0.33

Random Walker | Power Method

$$\begin{bmatrix} 0 & 0.33 & 0.33 & 0.33 \\ 0 & 0 & 0 & 1 \\ 0.5 & 0 & 0 & 0.5 \\ 0 & 0 & 1 & 0 \end{bmatrix}^{1}$$

Random Surfer model

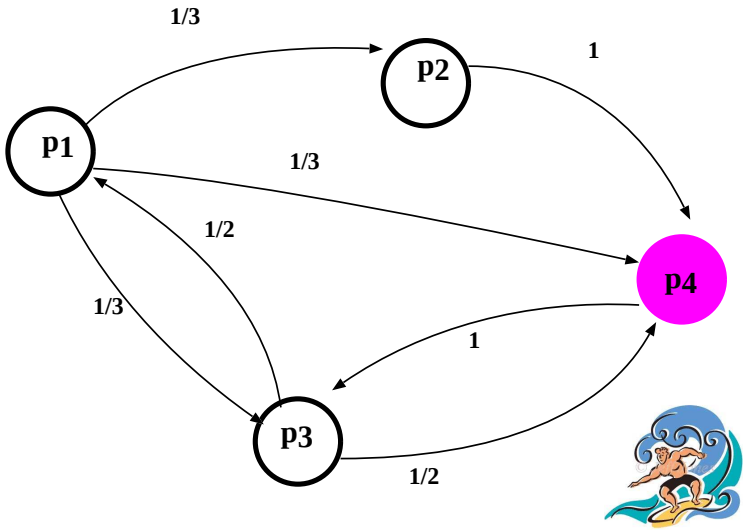


	# of Visits	Fraction of Visits	Probability of Visit
1	1	0.33	0.1667
2	0	0	0
3	1	0.33	0.3333
4	1	0.33	0.5

Random Walker — Power Method

$$\begin{bmatrix} 0 & 0.33 & 0.33 & 0.33 \\ 0 & 0 & 0 & 1 \\ 0.5 & 0 & 0 & 0.5 \\ 0 & 0 & 1 & 0 \end{bmatrix}^2$$

Random Surfer model

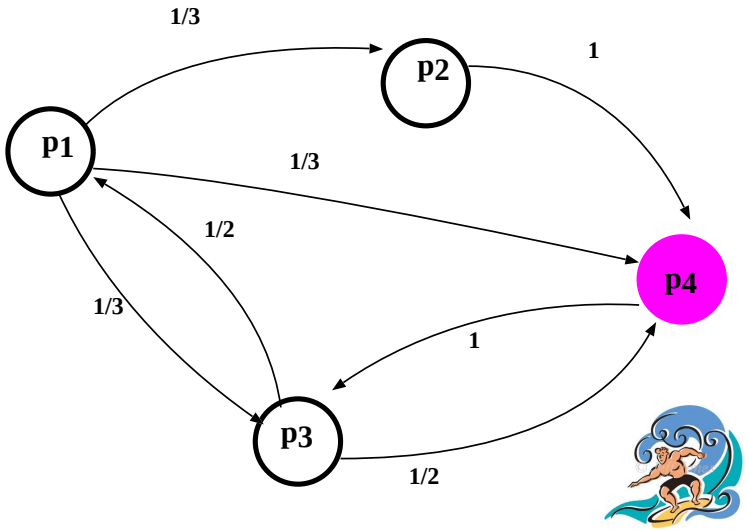


	# of Visits	Fraction of Visits	Probability of Visit
1	1	0.25	0.1667
2	0	0	0.0556
3	1	0.25	0.5556
4	2	0.5	0.2222

Random Walker — Power Method

$$\begin{bmatrix} 0 & 0.33 & 0.33 & 0.33 \\ 0 & 0 & 0 & 1 \\ 0.5 & 0 & 0 & 0.5 \\ 0 & 0 & 1 & 0 \end{bmatrix}^3$$

Random Surfer model



	# of Visits	Fraction of Visits	Probability of Visit
1	200	0.20	0.20
2	67	0.06	0.06
3	400	0.40	0.40
4	333	0.33	0.33

Random Walker — Power Method

$$\begin{bmatrix} 0 & 0.33 & 0.33 & 0.33 \\ 0 & 0 & 0 & 1 \\ 0.5 & 0 & 0 & 0.5 \\ 0 & 0 & 1 & 0 \end{bmatrix}^{1000}$$

Random Surfer model

- Let:
  - R_i be the ranking of a page i
  - n_i be the number of forward links out of page i
  - B_i be the set of pages that link to page i.
- The above ideas can be captured in the equation:

$$R_i = \sum_{j \epsilon B_i} R_j \frac{1}{n_j}$$

- This set of equations can be viewed as a “random web surfer” who visits pages and is equally likely to pick any of the forward links from the current page.
- These equations are the same as for a Markov chain!!!

Random Surfer model

- Let:
  - R_i be the ranking of a page i
  - n_i be the number of forward links out of page i
  - B_i be the set of pages that link to page i.
- The above ideas can be captured in the equation:

$$R_i = \sum_{j \epsilon B_i} R_j \frac{1}{n_j}$$

- This set of equations can be viewed as a “random web surfer” who visits pages and is equally likely to pick any of the forward links from the current page.
- These equations are the same as for a Markov chain!!!

Random Surfer model

- Let:
  - R_i be the ranking of a page i
  - n_i be the number of forward links out of page i
  - B_i be the set of pages that link to page i.
- The above ideas can be captured in the equation:

$$R_i = \sum_{j \epsilon B_i} R_j \frac{1}{n_j}$$

- This set of equations can be viewed as a “random web surfer” who visits pages and is equally likely to pick any of the forward links from the current page.
- These equations are the same as for a Markov chain!!!

Random Surfer model

- Let:
  - R_i be the ranking of a page i
  - n_i be the number of forward links out of page i
  - B_i be the set of pages that link to page i.
- The above ideas can be captured in the equation:

$$R_i = \sum_{j \epsilon B_i} R_j \frac{1}{n_j}$$

- This set of equations can be viewed as a “random web surfer” who visits pages and is equally likely to pick any of the forward links from the current page.
- These equations are the same as for a Markov chain!!!

Random Surfer model

- Let:
  - R_i be the ranking of a page i
  - n_i be the number of forward links out of page i
  - B_i be the set of pages that link to page i.
- The above ideas can be captured in the equation:

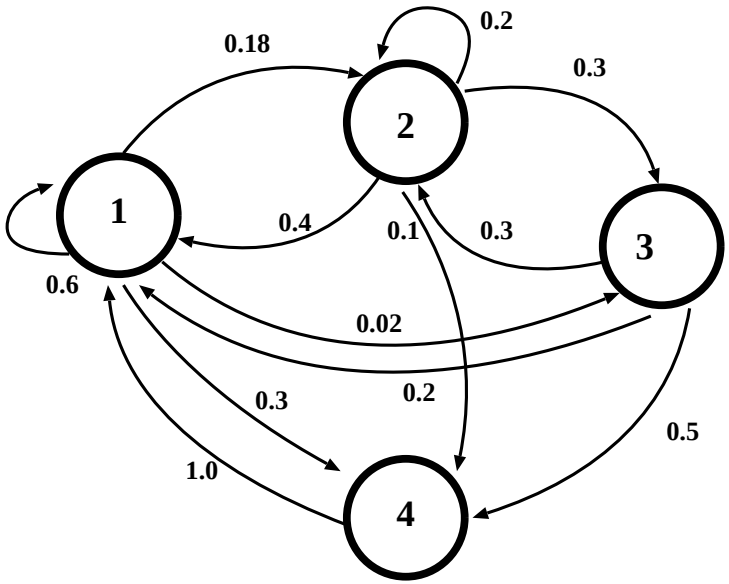
$$R_i = \sum_{j \epsilon B_i} R_j \frac{1}{n_j}$$

- This set of equations can be viewed as a “random web surfer” who visits pages and is equally likely to pick any of the forward links from the current page.
- These equations are the same as for a Markov chain!!!

Markov Chains

Everything you wanted to know about MCs and was afraid to ask

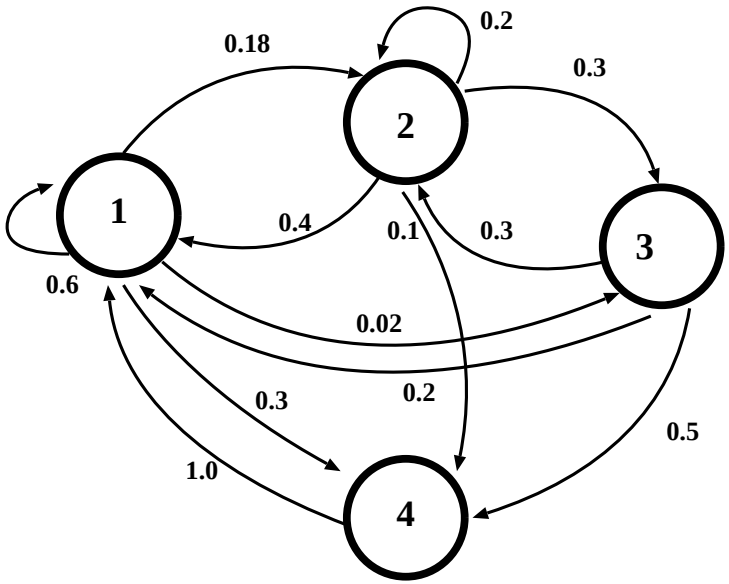
- state transition diagram
- directed graph.
- assign *transition probabilities* to arcs or *transition rates*.
- what a state represents? · · · Notion of state variables



Markov Chains

Everything you wanted to know about MCs and was afraid to ask

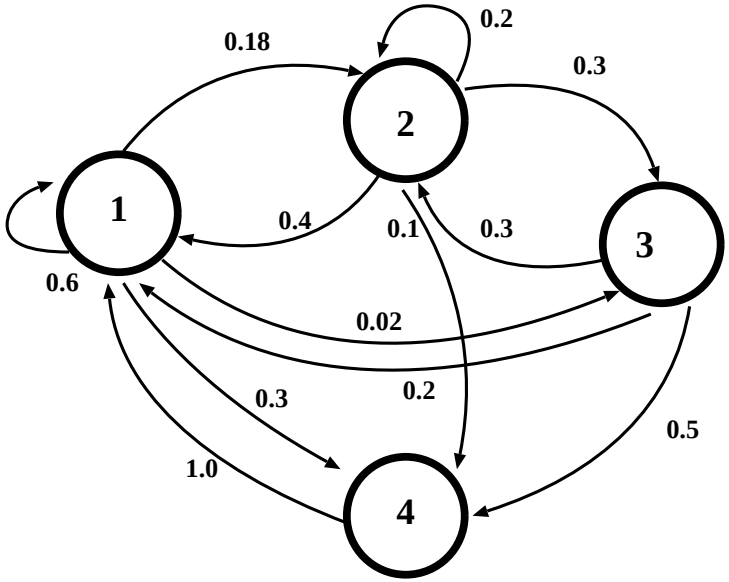
- state transition diagram
- directed graph.
- assign *transition probabilities* to arcs or *transition rates*.
- what a state represents? · · · Notion of state variables



Markov Chains

Everything you wanted to know about MCs and was afraid to ask

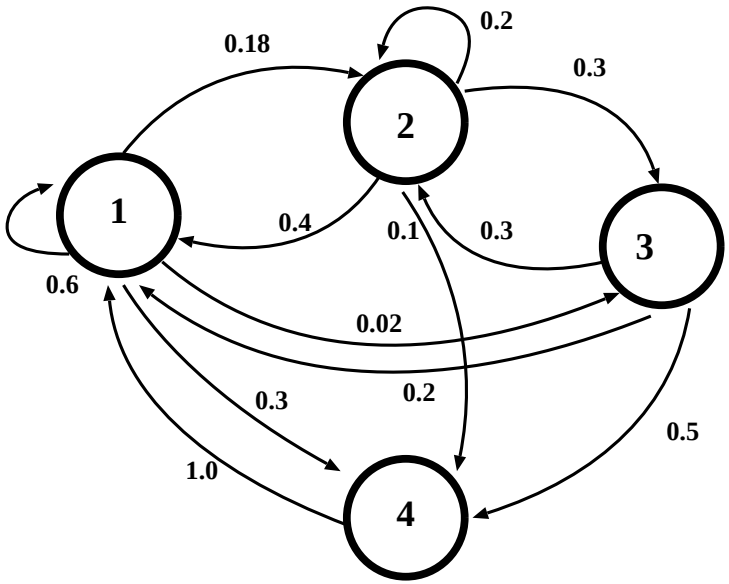
- state transition diagram
- directed graph.
- assign *transition probabilities* to arcs or *transition rates*.
- what a state represents? · · · Notion of state variables



Markov Chains

Everything you wanted to know about MCs and was afraid to ask

- state transition diagram
- directed graph.
- assign *transition probabilities* to arcs or *transition rates*.
- what a state represents? $\cdots$ Notion of state variables



Markov Chains
Measures of interest

- What can we calculate?
- fraction of time in a state
- fraction of visits to a state
- expected time to reach a set of states
- ⋮

Markov Chains
Measures of interest

- What can we calculate?
- fraction of time in a state
- fraction of visits to a state
- expected time to reach a set of states
- ⋮

Markov Chains
Measures of interest

- What can we calculate?
- fraction of time in a state
- fraction of visits to a state
- expected time to reach a set of states
- ⋮

Markov Chains
Measures of interest

- What can we calculate?
- fraction of time in a state
- fraction of visits to a state
- expected time to reach a set of states
-

Markov Chains
Measures of interest

- What can we calculate?
- fraction of time in a state
- fraction of visits to a state
- expected time to reach a set of states
- ⋮

Simple Example

- President of a large corporation travels every friday from one branch of the company to the other, and stays in that branch for the whole week.
- each branch is located in a different city. (eg. SP, Rio, MG, Brasilia, São Luis, Belem, Salvador, PA, Manaus)
- President only choses one particular company and choses the next city to visit if there is a non-stop flight from the current city to the next.
- If there are more than one choice for the next city, she chooses the next city at random.
- you know that the President has been doing this for a long time but you do not know whwere the President curently is.
- you want to meet the President, but you must be at the city she is in to succeed. Question: which city do you choose?

Simple Example

- President of a large corporation travels every friday from one branch of the company to the other, and stays in that branch for the whole week.
- each branch is located in a different city. (eg. SP, Rio, MG, Brasilia, São Luis, Belem, Salvador, PA, Manaus)
- President only choses one particular company and choses the next city to visit if there is a non-stop flight from the current city to the next.
- If there are more than one choice for the next city, she chooses the next city at random.
- you know that the President has been doing this for a long time but you do not know whwere the President curently is.
- you want to meet the President, but you must be at the city she is in to succeed. Question: which city do you choose?

Simple Example

- President of a large corporation travels every friday from one branch of the company to the other, and stays in that branch for the whole week.
- each branch is located in a different city. (eg. SP, Rio, MG, Brasilia, São Luis, Belem, Salvador, PA, Manaus)
- President only choses one particular company and choses the next city to visit if there is a non-stop flight from the current city to the next.
- If there are more than one choice for the next city, she chooses the next city at random.
- you know that the President has been doing this for a long time but you do not know whwere the President curently is.
- you want to meet the President, but you must be at the city she is in to succeed. Question: which city do you choose?

Simple Example

- President of a large corporation travels every friday from one branch of the company to the other, and stays in that branch for the whole week.
- each branch is located in a different city. (eg. SP, Rio, MG, Brasilia, São Luis, Belem, Salvador, PA, Manaus)
- President only choses one particular company and choses the next city to visit if there is a non-stop flight from the current city to the next.
- If there are more than one choice for the next city, she chooses the next city at random.
- you know that the President has been doing this for a long time but you do not know whwere the President curently is.
- you want to meet the President, but you must be at the city she is in to succeed. Question: which city do you choose?

Simple Example

- President of a large corporation travels every friday from one branch of the company to the other, and stays in that branch for the whole week.
- each branch is located in a different city. (eg. SP, Rio, MG, Brasilia, São Luis, Belem, Salvador, PA, Manaus)
- President only choses one particular company and choses the next city to visit if there is a non-stop flight from the current city to the next.
- If there are more than one choice for the next city, she chooses the next city at random.
- you know that the President has been doing this for a long time but you do not know whwere the President curently is.
- you want to meet the President, but you must be at the city she is in to succeed. Question: which city do you choose?

Simple Example

- President of a large corporation travels every friday from one branch of the company to the other, and stays in that branch for the whole week.
- each branch is located in a different city. (eg. SP, Rio, MG, Brasilia, São Luis, Belem, Salvador, PA, Manaus)
- President only choses one particular company and choses the next city to visit if there is a non-stop flight from the current city to the next.
- If there are more than one choice for the next city, she chooses the next city at random.
- you know that the President has been doing this for a long time but you do not know whwere the President curently is.
- you want to meet the President, but you must be at the city she is in to succeed. Question: which city do you choose?

Page Ranking: part 2

- Random surfer model: a large Discrete Time Markov Chain in which web pages are the states
- Google page rank is equivalent to the stationary state probabilities of this Markov Chain.
- The page ranks are proportional to the state probabilities.
- Solution: Power method (iterative method)

Page Ranking: part 2

- Random surfer model: a large Discrete Time Markov Chain in which web pages are the states
- Google page rank is equivalent to the stationary state probabilities of this Markov Chain.
- The page ranks are proportional to the state probabilities.
- Solution: Power method (iterative method)

Page Ranking: part 2

- Random surfer model: a large Discrete Time Markov Chain in which web pages are the states
- Google page rank is equivalent to the stationary state probabilities of this Markov Chain.
- The page ranks are proportional to the state probabilities.
- Solution: Power method (iterative method)

Page Ranking: part 2

- Random surfer model: a large Discrete Time Markov Chain in which web pages are the states
- Google page rank is equivalent to the stationary state probabilities of this Markov Chain.
- The page ranks are proportional to the state probabilities.
- Solution: Power method (iterative method)

Problems

- How to guarantee convergency?
- The web is too huge for it to be practical to crawl all web pages. So...
- The underlying Markov chain should be truncated...
- What to do with dangling nodes?
- dangling node: do not point to any page (pdf files, image files, etc)

Problems

- How to guarantee convergency?
- The web is too huge for it to be practical to crawl all web pages. So...
- The underlying Markov chain should be truncated...
- What to do with dangling nodes?
- dangling node: do not point to any page (pdf files, image files, etc)

Problems

- How to guarantee convergency?
- The web is too huge for it to be practical to crawl all web pages. So...
- The underlying Markov chain should be truncated...
- What to do with dangling nodes?
- dangling node: do not point to any page (pdf files, image files, etc)

Problems

- How to guarantee convergency?
- The web is too huge for it to be practical to crawl all web pages. So...
- The underlying Markov chain should be truncated...
- What to do with dangling nodes?
- dangling node: do not point to any page (pdf files, image files, etc)

Problems

- How to guarantee convergency?
- The web is too huge for it to be practical to crawl all web pages. So...
- The underlying Markov chain should be truncated...
- What to do with dangling nodes?
- dangling node: do not point to any page (pdf files, image files, etc)

Possible solution

- After entering a dangling node, the random surfer can hyperlink to any page. (equal probability)
- But other problems exist.
- For instance the Markov chain may not be aperiodic.

Possible solution

- After entering a dangling node, the random surfer can hyperlink to any page. (equal probability)
- But other problems exist.
- For instance the Markov chain may not be aperiodic.

Possible solution

- After entering a dangling node, the random surfer can hyperlink to any page. (equal probability)
- But other problems exist.
- For instance the Markov chain may not be aperiodic.

Problems

- Truncated model: we would like to have some idea of what portion of important pages are represented in this truncated model.
- One proposed measure: RankMass: defined as $\sum_{i=1}^{N} r_i$. This is a measure of the total page rank values captured in the truncated model.
- Problem: how to estimate or bound the RankMass for the portion of the web that has been crawled to form the truncated model?

Problems

- Truncated model: we would like to have some idea of what portion of important pages are represented in this truncated model.
- One proposed measure: RankMass: defined as $\sum_{i=1}^{N} r_i$. This is a measure of the total page rank values captured in the truncated model.
- Problem: how to estimate or bound the RankMass for the portion of the web that has been crawled to form the truncated model?

Problems

- Truncated model: we would like to have some idea of what portion of important pages are represented in this truncated model.
- One proposed measure: RankMass: defined as $\sum_{i=1}^{N} r_i$. This is a measure of the total page rank values captured in the truncated model.
- Problem: how to estimate or bound the RankMass for the portion of the web that has been crawled to form the truncated model?

Problems

- Needs assumptions since the undiscovered (not crawled) part of the web is totally unknown.
- That is, how to estimate the RankMass?

Problems

- Needs assumptions since the undiscovered (not crawled) part of the web is totally unknown.
- That is, how to estimate the RankMass?

Problems

- Needs assumptions since the undiscovered (not crawled) part of the web is totally unknown.
- That is, how to estimate the RankMass?

Solutions

- Create a new state that represents all the pages that have not been crawled. The dangling links will all be supposed to link to this page (state)
- Assumption: there is a set of trusted web pages.
  - with probability *d* the random surfer does not follow the hyperlink. Instead it jumps to one of the trusted pages (with equal probability)

Solutions

- Create a new state that represents all the pages that have not been crawled. The dangling links will all be supposed to link to this page (state)
- Assumption: there is a set of trusted web pages.
  - with probability *d* the random surfer does not follow the hyperlink. Instead it jumps to one of the trusted pages (with equal probability)

Solutions

- Let:
  - R be the vector of page ranks for the entire web
  - T be a vector which has all zero's except for non-zero values in positions corresponding to the trusted pages. In those locations the value is $\frac{1}{n_T}$,
  - $n_T = |T|$ is the cardinality of the set of trusted pages.
- Then... Google's equation

Solutions

- Let:
  - R be the vector of page ranks for the entire web
  - T be a vector which has all zero's except for non-zero values in positions corresponding to the trusted pages. In those locations the value is $\frac{1}{n_T}$,
  - $n_T = |T|$ is the cardinality of the set of trusted pages.
- Then... Google's equation

Solutions

- Google's equation

$$r_i = d \sum_{j \in \mathcal{I}_i} r_j \frac{1}{c_j} + (1-d)T[i] \quad \forall i$$

- Heinrich Hertz

  One cannot escape the feeling that these mathematical formulas have an independent existence and an intelligence of their own, that they are wiser than we are, wiser even than their discoverers, that we get more out of them than was originally put into them.

  (regarding Maxwell's equation)

Solutions

- We can bound the RankMass
- Use aggregation/disaggregation theory

Solutions

- We can bound the RankMass
- Use aggregation/disaggregation theory

Random Walks on Graphs

- Random surfer over web pages is a ramdom walk on a graph.
- Obtaining accurate statistics about web pages:
  - What percent of web pages are in the .com domain?
  - How many pages are indexed by a particular search engine?
  - What is the distribution of sizes, modification times, and content of web pages?
- Many other applications:
  - movement models for mobile computing
  - how closely related are 2 nodes in a graph?
  - obtaining relevance score between two nodes is one the fundamental problems in data mining
- In general a useful tool to calculate many measures of interest.

Random Walks on Graphs

- Random surfer over web pages is a ramdom walk on a graph.
- Obtaining accurate statistics about web pages:
  - What percent of web pages are in the .com domain?
  - How many pages are indexed by a particular search engine?
  - What is the distribution of sizes, modification times, and content of web pages?
- Many other applications:
  - movement models for mobile computing
  - how closely related are 2 nodes in a graph?
  - obtaining relevance score between two nodes is one the fundamental problems in data mining
- In general a useful tool to calculate many measures of interest.

Random Walks on Graphs

- Random surfer over web pages is a ramdom walk on a graph.
- Obtaining accurate statistics about web pages:
  - What percent of web pages are in the .com domain?
  - How many pages are indexed by a particular search engine?
  - What is the distribution of sizes, modification times, and content of web pages?
- Many other applications:
  - movement models for mobile computing
  - how closely related are 2 nodes in a graph?
  - obtaining relevance score between two nodes is one the fundamental problems in data mining
- In general a useful tool to calculate many measures of interest.

Random Walks on Graphs

- Random surfer over web pages is a ramdom walk on a graph.
- Obtaining accurate statistics about web pages:
  - What percent of web pages are in the .com domain?
  - How many pages are indexed by a particular search engine?
  - What is the distribution of sizes, modification times, and content of web pages?
- Many other applications:
  - movement models for mobile computing
  - how closely related are 2 nodes in a graph?
  - obtaining relevance score between two nodes is one the fundamental problems in data mining
- In general a useful tool to calculate many measures of interest.

Random Walks on Graphs

- Random surfer over web pages is a ramdom walk on a graph.
- Obtaining accurate statistics about web pages:
  - What percent of web pages are in the .com domain?
  - How many pages are indexed by a particular search engine?
  - What is the distribution of sizes, modification times, and content of web pages?
- Many other applications:
  - movement models for mobile computing
  - how closely related are 2 nodes in a graph?
  - obtaining relevance score between two nodes is one the fundamental problems in data mining
- In general a useful tool to calculate many measures of interest.

Random Walks on Graphs

- Random surfer over web pages is a ramdom walk on a graph.
- Obtaining accurate statistics about web pages:
  - What percent of web pages are in the .com domain?
  - How many pages are indexed by a particular search engine?
  - What is the distribution of sizes, modification times, and content of web pages?
- Many other applications:
  - movement models for mobile computing
  - how closely related are 2 nodes in a graph?
  - obtaining relevance score between two nodes is one the fundamental problems in data mining
- In general a useful tool to calculate many measures of interest.

Random Walks on Graphs

- Abstract model: graph model, probability associated with edges
- At any time there is a current vertex (state)
- The next vertex is chosen at random among the neighbors of the current vertex
- Clearly this defines a Markov chain

Random Walks on Graphs

- Abstract model: graph model, probability associated with edges
- At any time there is a current vertex (state)
- The next vertex is chosen at random among the neighbors of the current vertex
- Clearly this defines a Markov chain

Random Walks on Graphs

- Abstract model: graph model, probability associated with edges
- At any time there is a current vertex (state)
- The next vertex is chosen at random among the neighbors of the current vertex
- Clearly this defines a Markov chain

Random Walks on Graphs

- Abstract model: graph model, probability associated with edges
- At any time there is a current vertex (state)
- The next vertex is chosen at random among the neighbors of the current vertex
- Clearly this defines a Markov chain

Random Walks on Graphs

- a class of random walks can be identified as a special type of MC.
- Can use theory of reversible MCs.

Random Walks on Graphs

- a class of random walks can be identified as a special type of MC.
- Can use theory of reversible MCs.

Some Example Applications
path exists?

- For a graph G, determine if there is a path from verticies s to d.
- Proposal:
  - start a random walk at s, run for up to N steps and return path if d is reached, or else return *no path*.
  - Note: this is a *randomized algorithm*
- *No path* can be incorrect
- Intuitively, as N grows, the probability of not finding a path if one exists can be made small
- What can be said about the relationship of N and the probability of not finding a path?
- Use MC theory, time reversibility, etc.

Some Example Applications
path exists?

- For a graph G, determine if there is a path from verticies s to d.
- Proposal:
  - start a random walk at s, run for up to N steps and return path if d is reached, or else return *no path*.
  - Note: this is a *randomized algorithm*
- *No path* can be incorrect
- Intuitively, as N grows, the probability of not finding a path if one exists can be made small
- What can be said about the relationship of N and the probability of not finding a path?
- Use MC theory, time reversibility, etc.

Some Example Applications
path exists?

- For a graph G, determine if there is a path from verticies s to d.
- Proposal:
  - start a random walk at s, run for up to N steps and return path if d is reached, or else return *no path*.
  - Note: this is a *randomized algorithm*
- *No path* can be incorrect
- Intuitively, as N grows, the probability of not finding a path if one exists can be made small
- What can be said about the relationship of N and the probability of not finding a path?
- Use MC theory, time reversibility, etc.

Some Example Applications
path exists?

- For a graph G, determine if there is a path from verticies s to d.
- Proposal:
  - start a random walk at s, run for up to N steps and return path if d is reached, or else return *no path*.
  - Note: this is a *randomized algorithm*
- *No path* can be incorrect
- Intuitively, as N grows, the probability of not finding a path if one exists can be made small
- What can be said about the relationship of N and the probability of not finding a path?
- Use MC theory, time reversibility, etc.

Some Example Applications
path exists?

- For a graph G, determine if there is a path from verticies s to d.
- Proposal:
  - start a random walk at s, run for up to N steps and return path if d is reached, or else return *no path*.
  - Note: this is a *randomized algorithm*
- *No path* can be incorrect
- Intuitively, as N grows, the probability of not finding a path if one exists can be made small
- What can be said about the relationship of N and the probability of not finding a path?
- Use MC theory, time reversibility, etc.

Some Example Applications
path exists?

- For a graph G, determine if there is a path from verticies s to d.
- Proposal:
  - start a random walk at s, run for up to N steps and return path if d is reached, or else return *no path*.
  - Note: this is a *randomized algorithm*
- *No path* can be incorrect
- Intuitively, as N grows, the probability of not finding a path if one exists can be made small
- What can be said about the relationship of N and the probability of not finding a path?
- Use MC theory, time reversibility, etc.

Some Example Applications
social network

- Consider a social network graph. Links are friend relationships
- Objective: generate uniformly random samples to estimate the mean number of some quantities such as age
- Idea:
  - Let M be the maximum number of friends anyone has
  - simulate a random walk for some large number of steps. (until stationary state probabilities are approximately reached)
  - take a sample and do it again
  - Theory: you get uniform random samples.
- Issue: how many steps do you have to take to get close enough to stationary state probabilities?
- Issue: what about a dynamic graph?

Some Example Applications
social network

- Consider a social network graph. Links are friend relationships
- Objective: generate uniformly random samples to estimate the mean number of some quantities such as age
- Idea:
  - Let M be the maximum number of friends anyone has
  - simulate a random walk for some large number of steps. (until stationary state probabilities are approximately reached)
  - take a sample and do it again
  - Theory: you get uniform random samples.
- Issue: how many steps do you have to take to get close enough to stationary state probabilities?
- Issue: what about a dynamic graph?

Some Example Applications
social network

- Consider a social network graph. Links are friend relationships
- Objective: generate uniformly random samples to estimate the mean number of some quantities such as age
- Idea:
  - Let M be the maximum number of friends anyone has
  - simulate a random walk for some large number of steps. (until stationary state probabilities are approximately reached)
  - take a sample and do it again
  - Theory: you get uniform random samples.
- Issue: how many steps do you have to take to get close enough to stationary state probabilities?
- Issue: what about a dynamic graph?

Some Example Applications
social network

- Consider a social network graph. Links are friend relationships
- Objective: generate uniformly random samples to estimate the mean number of some quantities such as age
- Idea:
  - Let M be the maximum number of friends anyone has
  - simulate a random walk for some large number of steps. (until stationary state probabilities are approximately reached)
  - take a sample and do it again
  - Theory: you get uniform random samples.
- Issue: how many steps do you have to take to get close enough to stationary state probabilities?
- Issue: what about a dynamic graph?

Some Example Applications
social network

- Consider a social network graph. Links are friend relationships
- Objective: generate uniformly random samples to estimate the mean number of some quantities such as age
- Idea:
  - Let M be the maximum number of friends anyone has
  - simulate a random walk for some large number of steps. (until stationary state probabilities are approximately reached)
  - take a sample and do it again
  - Theory: you get uniform random samples.
- Issue: how many steps do you have to take to get close enough to stationary state probabilities?
- Issue: what about a dynamic graph?

Some Example Applications

Markov Chain Monte Carlo Method

- Consider a random variable X over a space $\mathcal{S}$
- Suppose we know the density for X.
- In a MCMC method one can generate random samples of X using random walks over $\mathcal{S}$.

Some Example Applications

Markov Chain Monte Carlo Method

- Consider a random variable X over a space $\mathcal{S}$
- Suppose we know the density for X.
- In a MCMC method one can generate random samples of X using random walks over $\mathcal{S}$.

Some Example Applications

Markov Chain Monte Carlo Method

- Consider a random variable X over a space $\mathcal{S}$
- Suppose we know the density for X.
- In a MCMC method one can generate random samples of X using random walks over $\mathcal{S}$.

Some Example Applications

Markov Chain Monte Carlo Method

- Consider a random variable X over a space $\mathcal{S}$
- Suppose we know the density for X.
- In a MCMC method one can generate random samples of X using random walks over $\mathcal{S}$.

THANKS

Hope you cameback tomorrow for more interesting problems