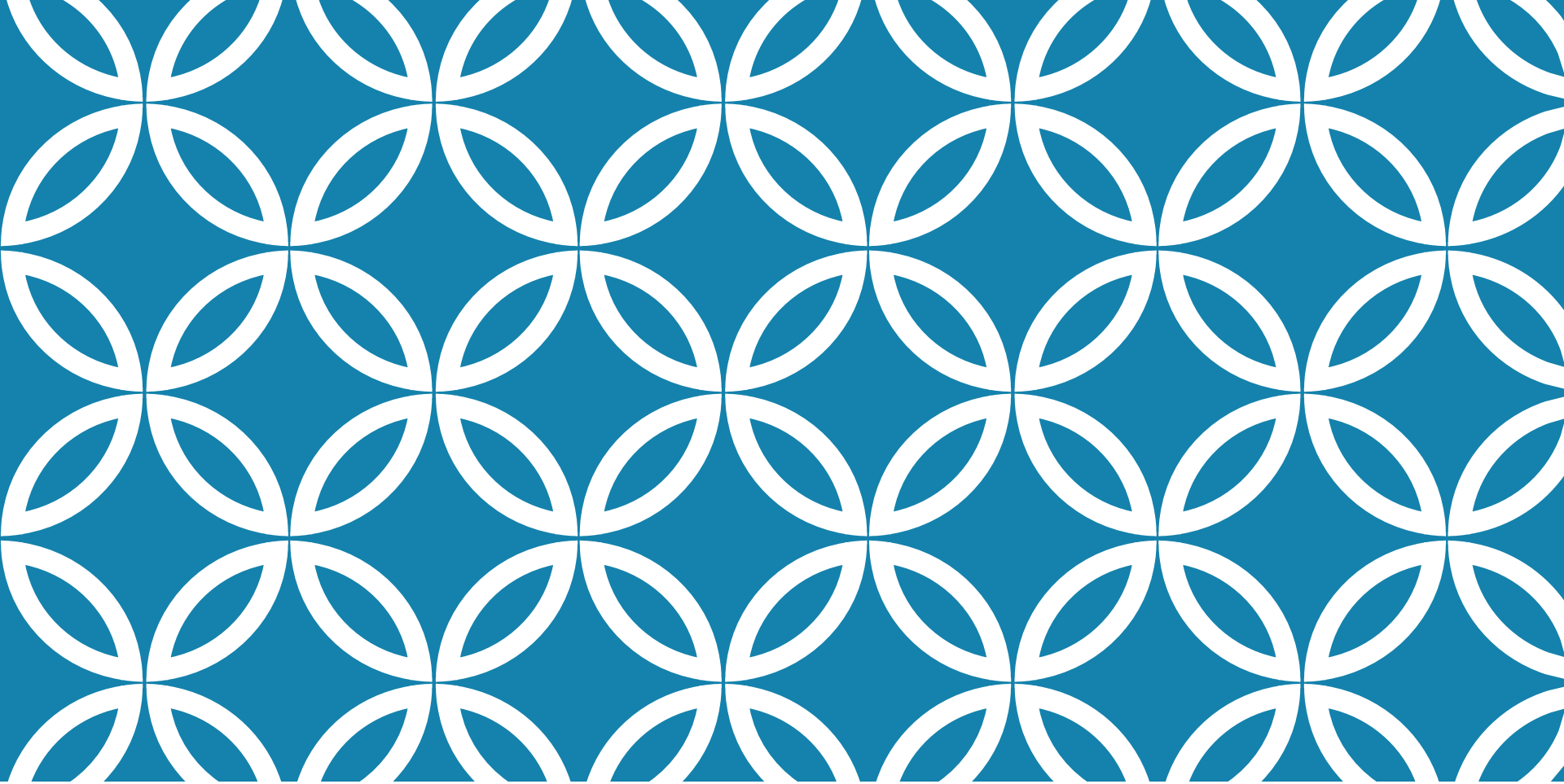


# ANÁLISE DE LINKS E BUSCA NA WEB

Redes Sociais e Econômicas

Prof. André Vignatti

# ANÁLISE DE LINKS E BUSCA NA WEB

quando você digita **"UFPR"** no Google, o primeiro resultado é www.ufpr.br

- como o **Google** sabe que essa é a melhor resposta a mostrar?

métodos de **busca na Web** usam a **informação contida na própria rede**, e não uma informação externa

antes de apresentar alguns métodos de busca na web, vamos entender porque é um problema difícil

# BUSCA NA WEB: O PROBLEMA DO RANKING

**recuperação de informação**: buscar documentos, informações dentro de documentos e na WWW

**problemas** com **palavras-chave**:

- **sinonímia**: várias formas de dizer a mesma coisa
  - ex: mandioquinha e batata salsa
- **polissemia**: vários significados para o mesmo termo
  - ex: **jaguar**: carro, animal, videogame, time de futebol americano, SO da Apple

# BUSCA NA WEB: O PROBLEMA DO RANKING

**problemas** relacionados à **escala e complexidade:**

- **anos 80 - busca feita em bibliotecas:** pessoas treinadas para fazer a busca sobre informações escritas por profissionais, usando vocabulário e estilo padronizados

- **anos 90 - chegada da Web:** qualquer novato ou expert pode produzir e buscar conteúdos

# BUSCA NA WEB: O PROBLEMA DO RANKING

**problemas** relacionados à **informação dinâmica**:

- as páginas da Web mudam constantemente
- em *11 de setembro de 2001*, muitos buscaram **"World Trade Center"** no Google
- na época, o Google coletava e indexava informações periodicamente, com **intervalos de semanas**
- hoje, máquinas de busca especializadas em "busca de notícias" foram incluídas no Google
    - mesmo assim, a integração não é perfeita
- **Twitter** preenche esses espaços entre conteúdo estático e de tempo real

# BUSCA NA WEB: O PROBLEMA DO RANKING

**problema**: **escassez X abundância**

- *antigamente*, o problema era escassez de informação: achar uma agulha do palheiro
- neste caso, a busca retornava informação sem sentido, o mais próximo do que ela havia conseguido
- *hoje,* o problema é abundância: uma busca retorna milhões de páginas relevantes, mas devemos saber **como "filtrar" e rankear** essa grande quantidade de informação

**resumindo**: a Web continua trazendo novos desafios para a recuperação de informação...

# ANÁLISE DE LINKS: VOTAÇÃO POR IN-LINKS

uma página por si só pode não ter muita informação para fazer o ranking

ao invés disso: **usar a Web inteira** para fazer o ranking

a ideia é usar a votação por **links entrantes (in-links):**

> podemos usar links para avaliar a **autoridade (a importância)** de uma página sobre um assunto

por exemplo, consultando **"UFPR"**: coleta páginas que contenham um **link com a palavra "UFPR"**, faz-se uma **"votação"** através das seus links para descobrir os mais proeminentes

# VOTAÇÃO POR IN-LINKS

algumas observações:

1. um link para “UFPR”, não significa necessariamente algo bom

   - pode ser **off-topic**, pode ser uma **crítica**, ou pode ser uma **propaganda**

   - mas é esperado que, na maioria, sejam links com um aval positivo

   - esse é um dos **argumentos que justifica** a análise por links

2. mesmo a estratégia simples de votação por in-links pode produzir resultados bons para **páginas populares**

   - talvez nem seja necessário analisar toda a Web, mas **somente uma amostra**

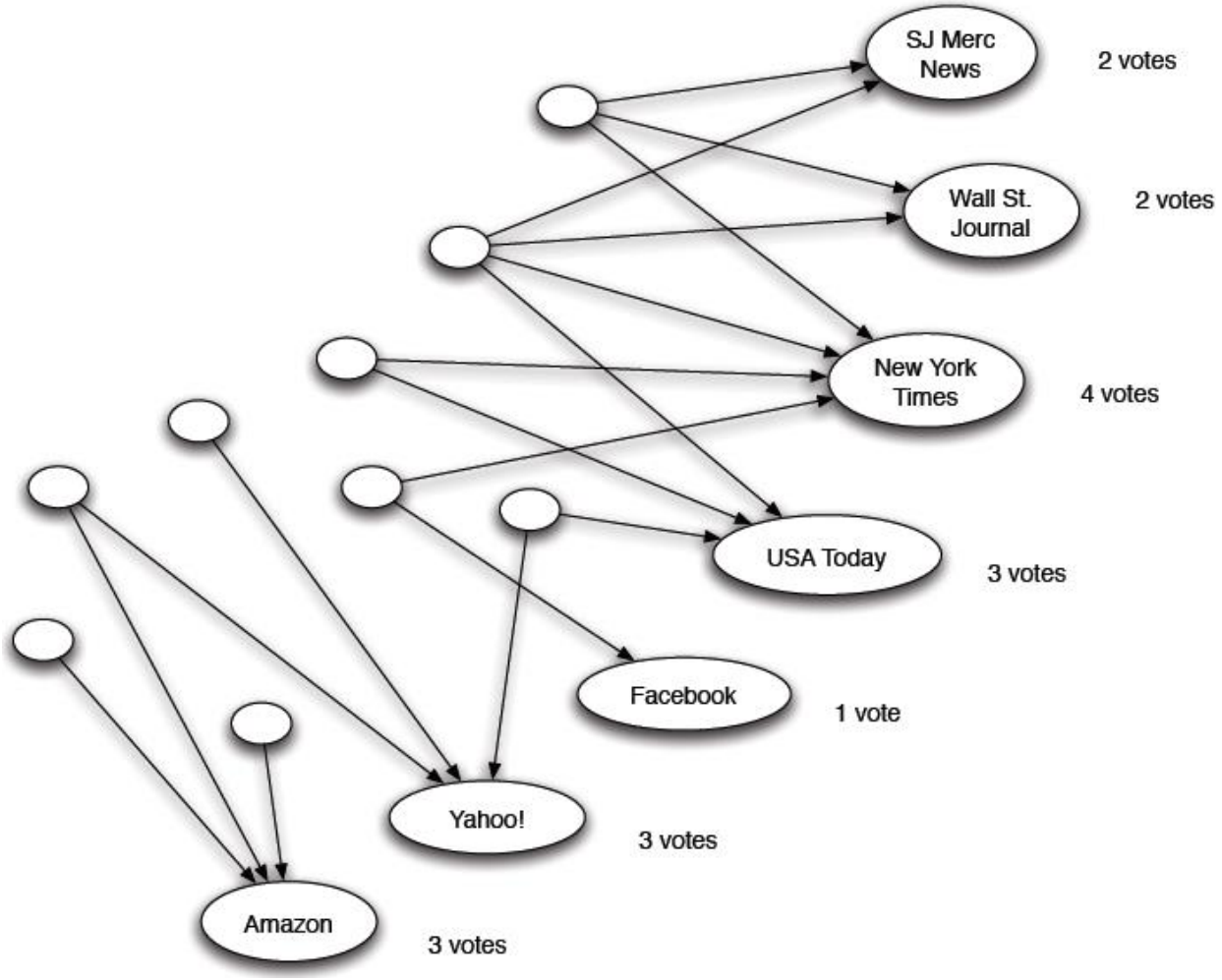
# A TÉCNICA DE LISTAS DE LINKS

considere a consulta **"newspapers"**

- ao contrário de **"UFPR"**, não há necessariamente uma única melhor resposta, há **vários jornais de destaque** na Web

- então, o resultado de uma votação por in-links terá **dois tipos de sites**: alguns **jornais de verdade**, e outros **sites genéricos** (MSN, Facebook, Amazon, Yahoo!, …)

# A TÉCNICA DE LISTAS DE LINKS



# A TÉCNICA DE LISTAS DE LINKS

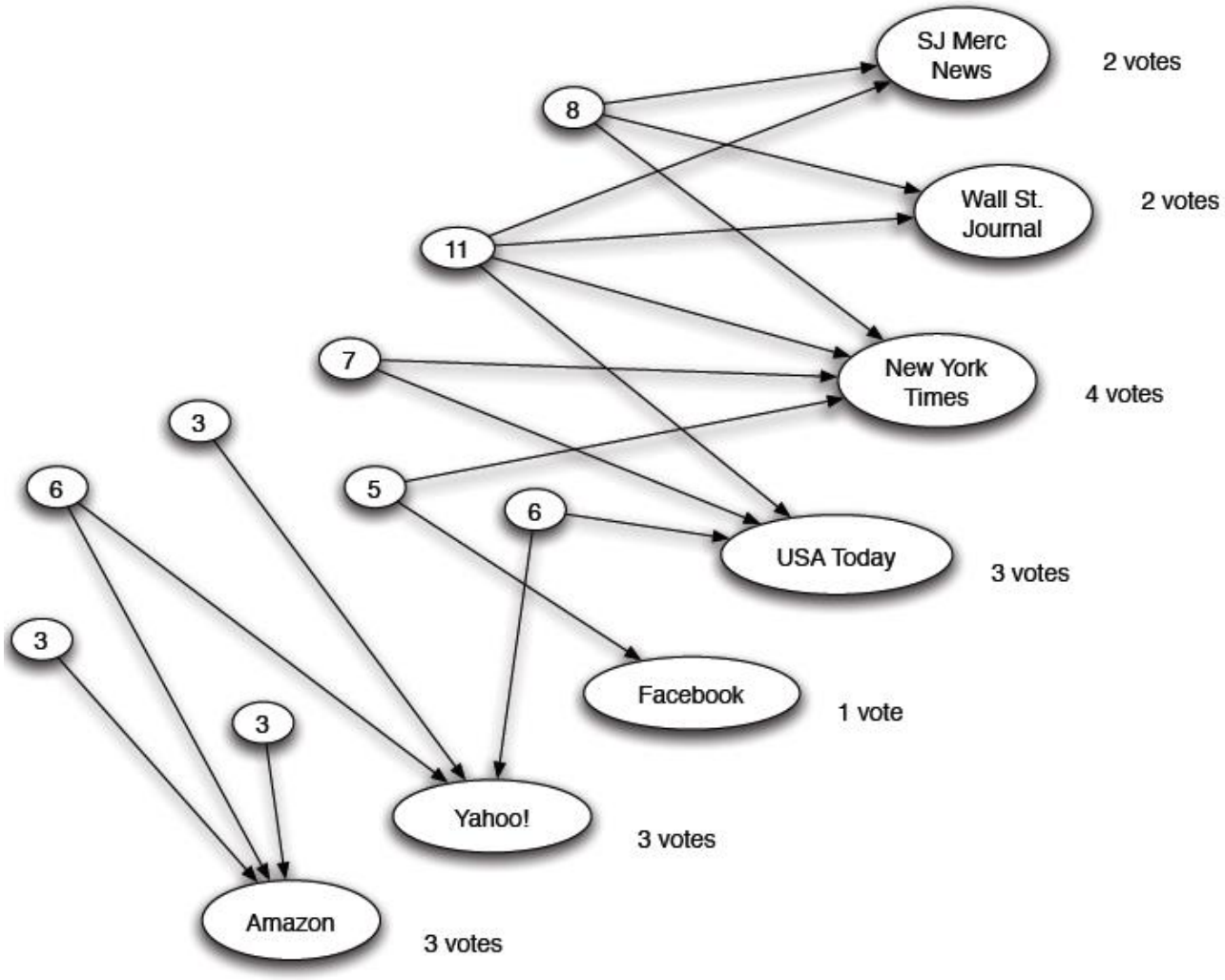
observando melhor, **ganhamos uma informação adicional**: páginas que **compilam listas de links** relevantes à busca

ou seja, muitas das páginas que "votaram", votaram nas páginas que receberam mais votos

assim, suspeitamos que essas páginas sejam **listas de "boas respostas"**

damos uma **nota para cada lista**: é a **soma dos votos da página que ela votou**

# A TÉCNICA DE LISTAS DE LINKS

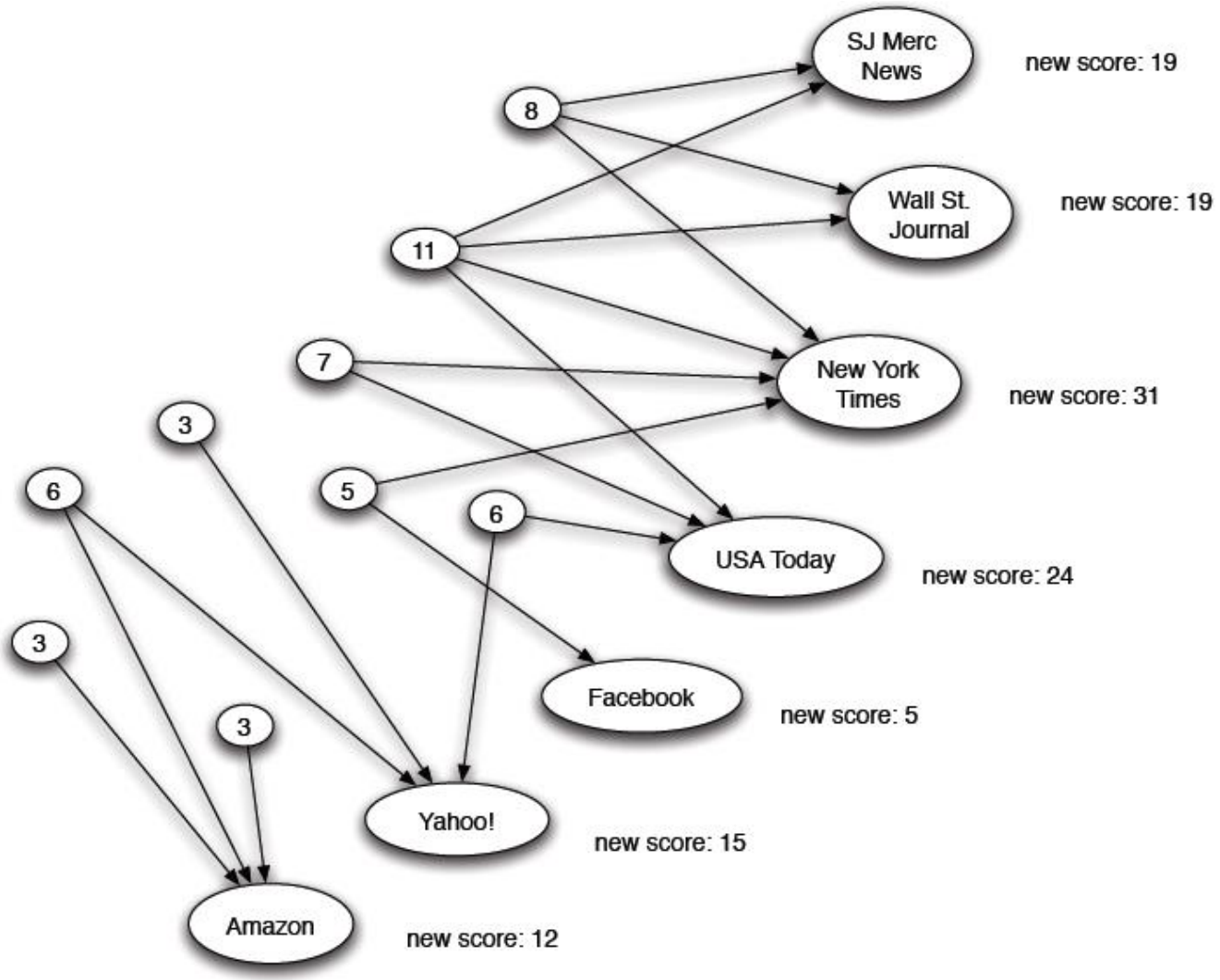


# A TÉCNICA DE LISTAS DE LINKS: MELHORIA REPETIDA

se as listas de links com boa pontuação tem uma melhor noção dos resultados bons, então devemos **dar um peso maior** a seus votos

assim, podemos **recalcular os votos**, mas dessa vez dando um **peso igual à soma das listas que votam**

# MELHORIA REPETIDA: EXEMPLO



# A TÉCNICA DE LISTAS DE LINKS: MELHORIA REPETIDA

agora temos votos melhores no lado direito da figura

podemos usar esses valores para **obter valores ainda mais refinados** para a qualidade das listas no lado esquerdo da figura

**melhoria repetida**: atualizar um lado de cada vez, várias vezes, **refinando cada vez mais** os valores

# HUBS E AUTORIDADES

as ideias anteriores **sugerem um método de ranking** na busca web, como vamos explicar

**autoridades**: as páginas que serão resposta da busca

**hubs**: as listas para a busca

para cada página $\boldsymbol{p}$ estimamos **dois valores**:

- $\boldsymbol{auth(p)}$: o valor como uma potencial autoridade
- $\boldsymbol{hub(p)}$: o valor como um potencial hub

# HUBS E AUTORIDADES

começamos céticos: $\boldsymbol{auth(p) = 1}$ e $\boldsymbol{hub(p) = 1}$

**regra de atualização da autoridade:** para cada página $\boldsymbol{p}$, atualizar $\boldsymbol{auth(p)}$ como sendo a soma dos valores de hub de todas as páginas que apontam para ela

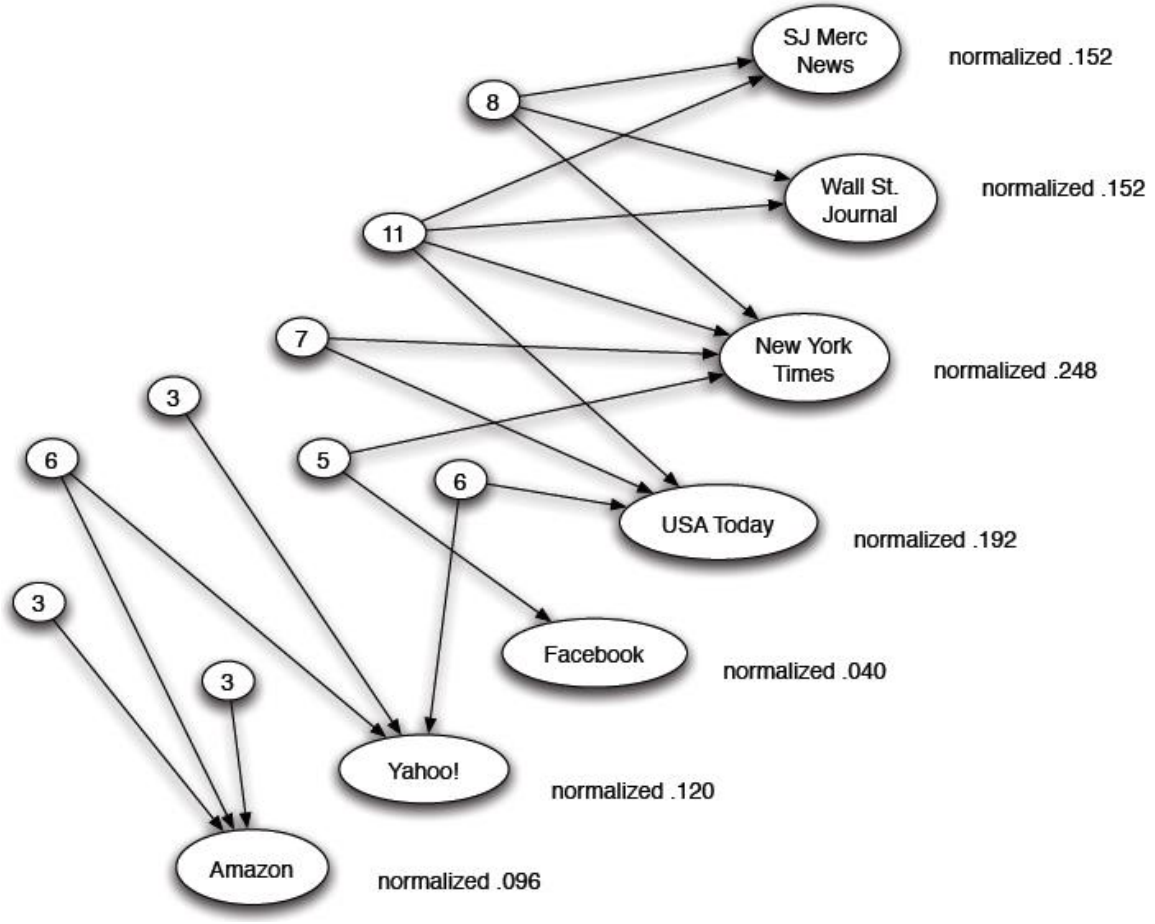
**regra de atualização da hub:** para cada página $\boldsymbol{p}$, atualizar $\boldsymbol{hub(p)}$ como sendo a soma dos valores de autoridade de todas as páginas que ela aponta

# HUBS E AUTORIDADES: MÉTODO BASEADO NA MELHORIA REPETIDA

1. comece com todos os valores (hubs e auth) = $\boldsymbol{1}$
2. escolha um número de passos $\boldsymbol{k}$
3. faça uma sequência de $\boldsymbol{k}$ atualizações de hubs e autoridades:
   - aplique a regra de atualização de autoridade para o conjunto atual de valores
   - aplique a regra de atualização do hub para o conjunto resultante de valores
4. normalize:
   - divida o valor de cada autoridade pela soma de todos valores de autoridades
   - divida o valor de cada hub pela soma de todos valores de hubs

# HUBS E AUTORIDADES: ETAPA DE NORMALIZAÇÃO

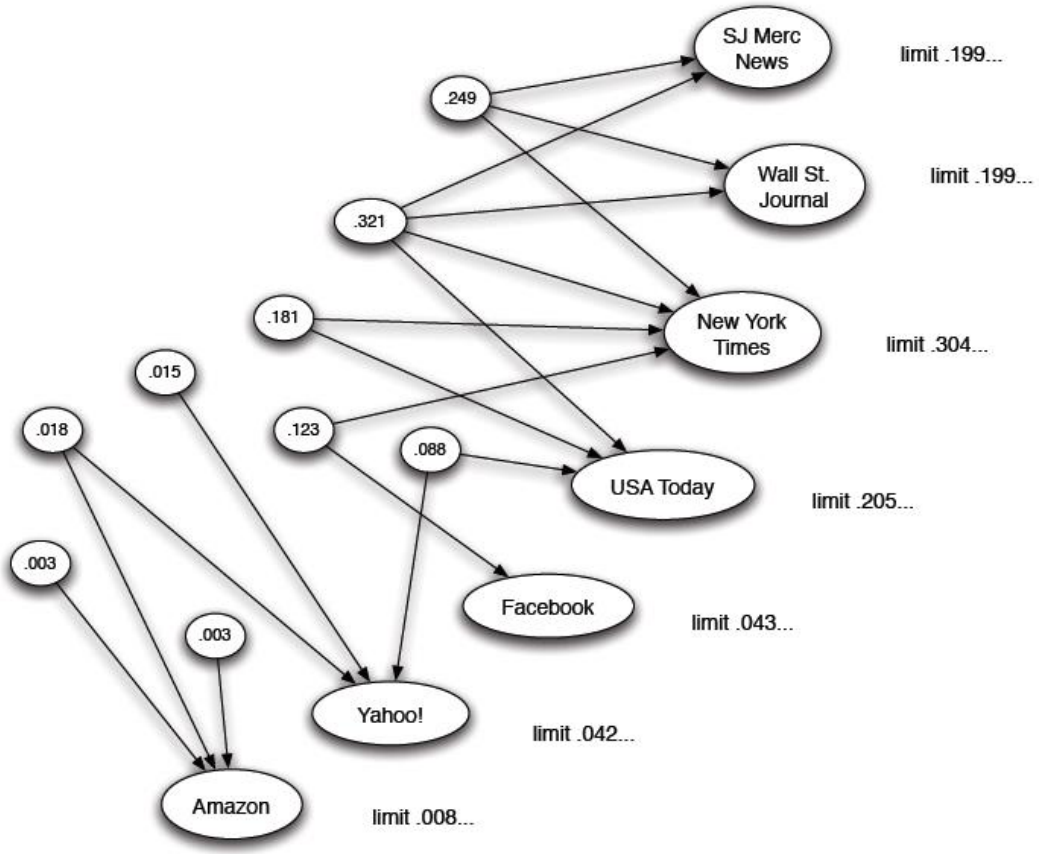
a normalização é feita pois queremos somente os **valores relativos**



# HUBS E AUTORIDADES: EQUILÍBRIO DOS VALORES

o que acontece quando $k$ (o número de passos) aumenta?

- se $k \to \infty$, os valores de hubs e autoridades **convergem para um limite**!



# HUBS E AUTORIDADES: EQUILÍBRIO DOS VALORES

**além disso**: podemos usar qualquer valor inicial (nós usamos 1), que os limites serão os mesmos

- ou seja, *os limites não dependem dos valores iniciais*

**conclusão**: os valores de hubs e autoridades são **uma propriedade intrínsica da estrutura** de links da rede

esse valores de limite refletem um equilíbrio entre hubs e autoridades: se aplicarmos as regras de atualização, o valor não mudará!

**demonstração formal**: Seção 14.6 (não veremos aqui)