

Link Analysis and Web Search

מבוסס על פרק 14

מוצג על ידי: **תמיר שאשא**

תאריך: 20 נובמבר 2014

הונחה על ידי: ד"ר **רונית רובינפלד**

על מה נדבר היום?

- What is the motivation and difficulties to rank pages?
- Different techniques to rank pages
- Page Ranking in practice

קצת מוטיבציה

מדוע צריך לדרג דפים?

מדוע זאת בעיה קשה לדרג דפים?

למה צריך להתייחס כאשר מדרגים דפים?

טכניקות לדירוג דפים

בהמשך נציג טכניקות שונות לדירוג דפים כמו:

•Link Analysis using Hubs and Authorities

•PageRank

טכניקה ראשונה

.Link Analysis using Hubs and Authorities

- Voting By In Links
- The Principle Of Repeated Improvement
- Hubs And Authorities

טכניקה ראשונה: נקודות חשובות

האם כאשר מילות החיפוש מופיעות פעמים רבות זהו סימן לתוצאה טובה?

חשיבות הדף היא **פונקציה** של הדף עצמו ושל דפים הקשורים אליו

Voting By In-Links

הרעיון :

דף הוא חשוב אם הרבה דפים
מצביעים עליו

Voting By In-Links

כיצד מדרגים דפים?

שלב ראשון

- קיבלנו דפים המכילים את מילות החיפוש.
- נאסוף דפים המצביעים לדפים ברשימת התוצאות.
- נספור את מספר ההצבעות של כל דף ברשימה.

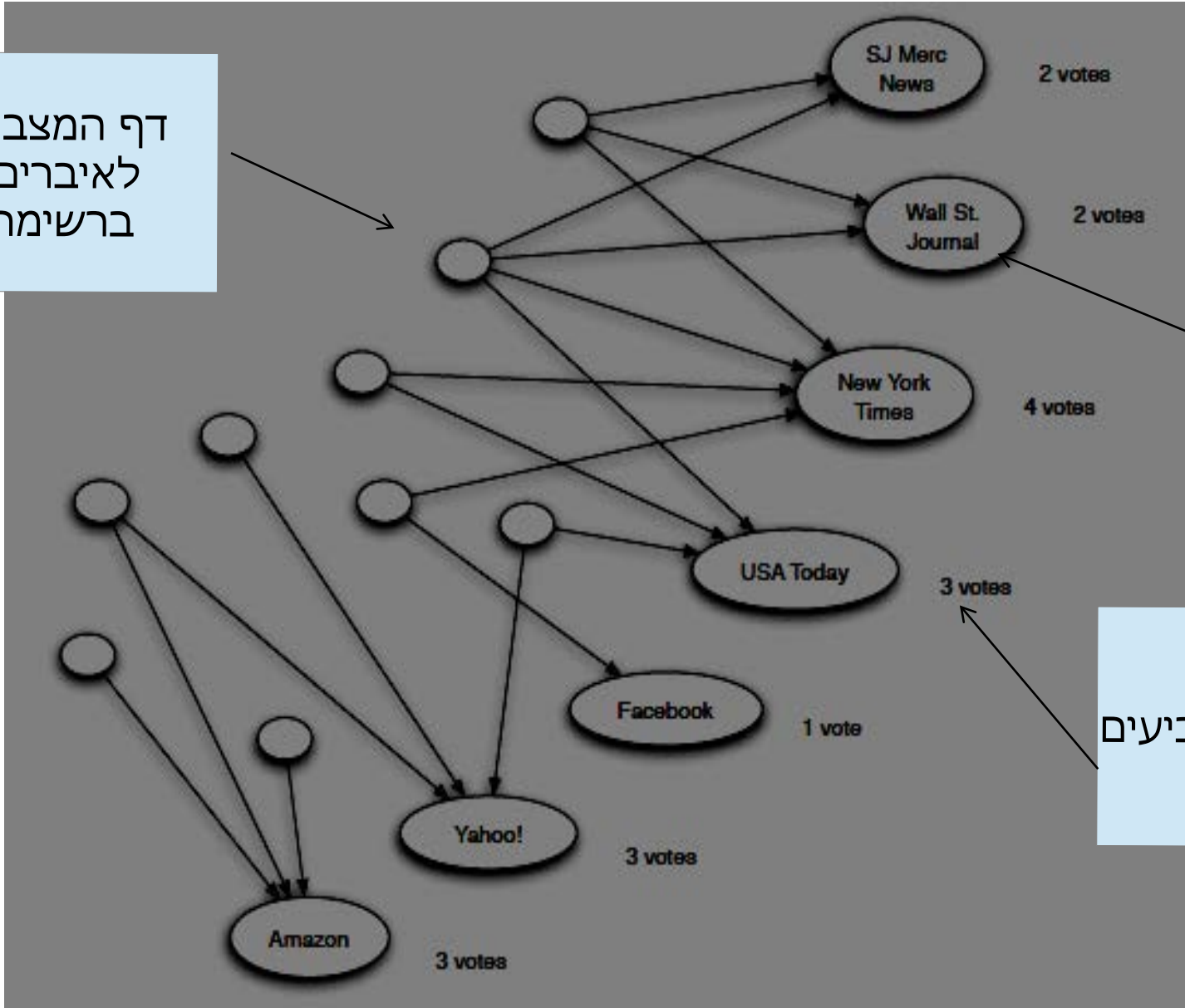


Voting By In-Links - Example

דף המצביע
לאיברים
ברשימה

איבר ברשימה
שהמילה
מופיעה בו

מספר המצביעים



Voting By In-Links

מסקנה

דפים שנראים **חשובים** קיבלו **מספר גבוה** של הצבעות



Voting By In-Links

נשים לב שבדוגמא שראינו, מבין הדפים שקיבלו את מספר ההצבעות הגבוה,

2. לא היו עיתונים -אלו דפים שיקבלו מספר הצבעות לא משנה מהי מילת החיפוש

A List Finding Technique

נשתמש במבנה הרשת בצורה חכמה יותר.

למצביעים שונים השפעה שונה.

דירוג מחדש של רשימת התוצאות.

A List Finding Technique

ישנם **מעט דפים** המצביעים לדפים בעלי מספר הצבעות גבוה

.ניתן **משקל גבוה** יותר לדפים מהסוג הנ"ל

.הערך יהיה **סכום ההצבעות** של הדפים ברשימה שהם **הצביעו** עליהם

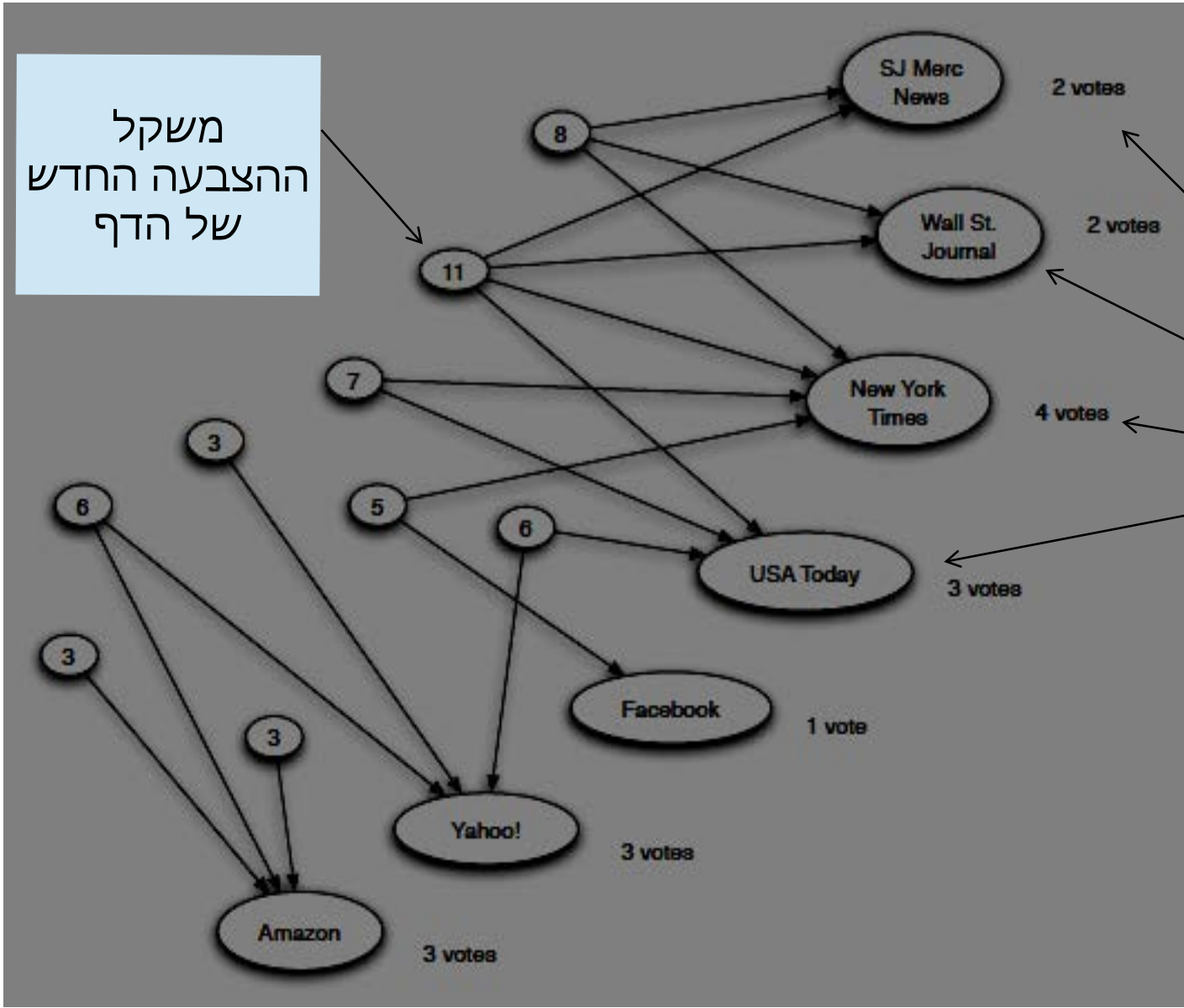


שלב שני

דוגמא

משקל
ההצבעה החדש
של הדף

מספר ההצבעות
שקיבלו הדפים
המוצבעים
על ידי הדף



A List Finding Technique

נעדכן את ערך ההצבעות בהתאם למשקל החדש
– ערך המשקל שמצאנו בשלב השני

דירוג החדש של כל דף ברשימה הוא **סכום**
המשקלי ההצבעה של הדפים **המצביעים** עליו

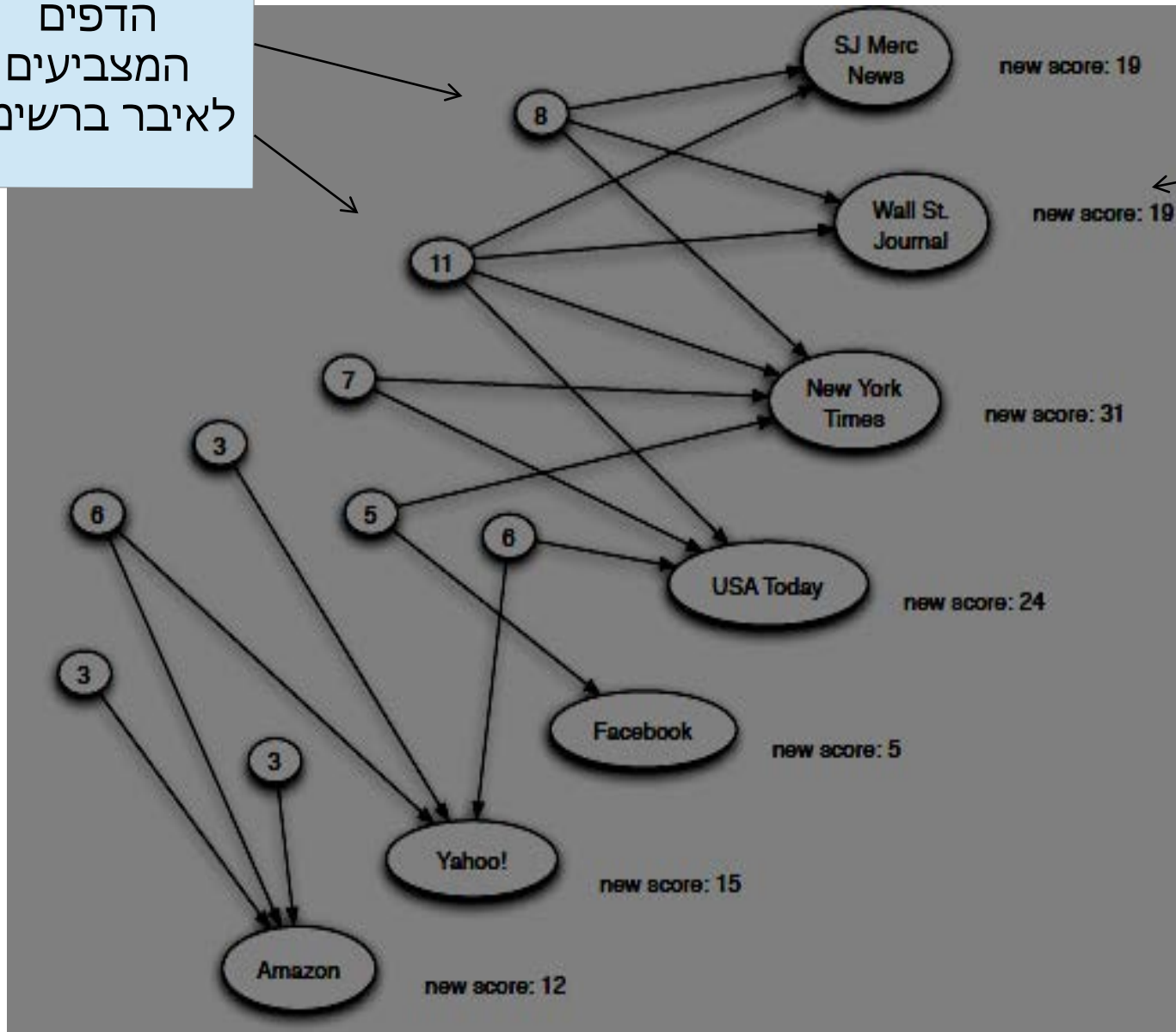


שלב שלישי

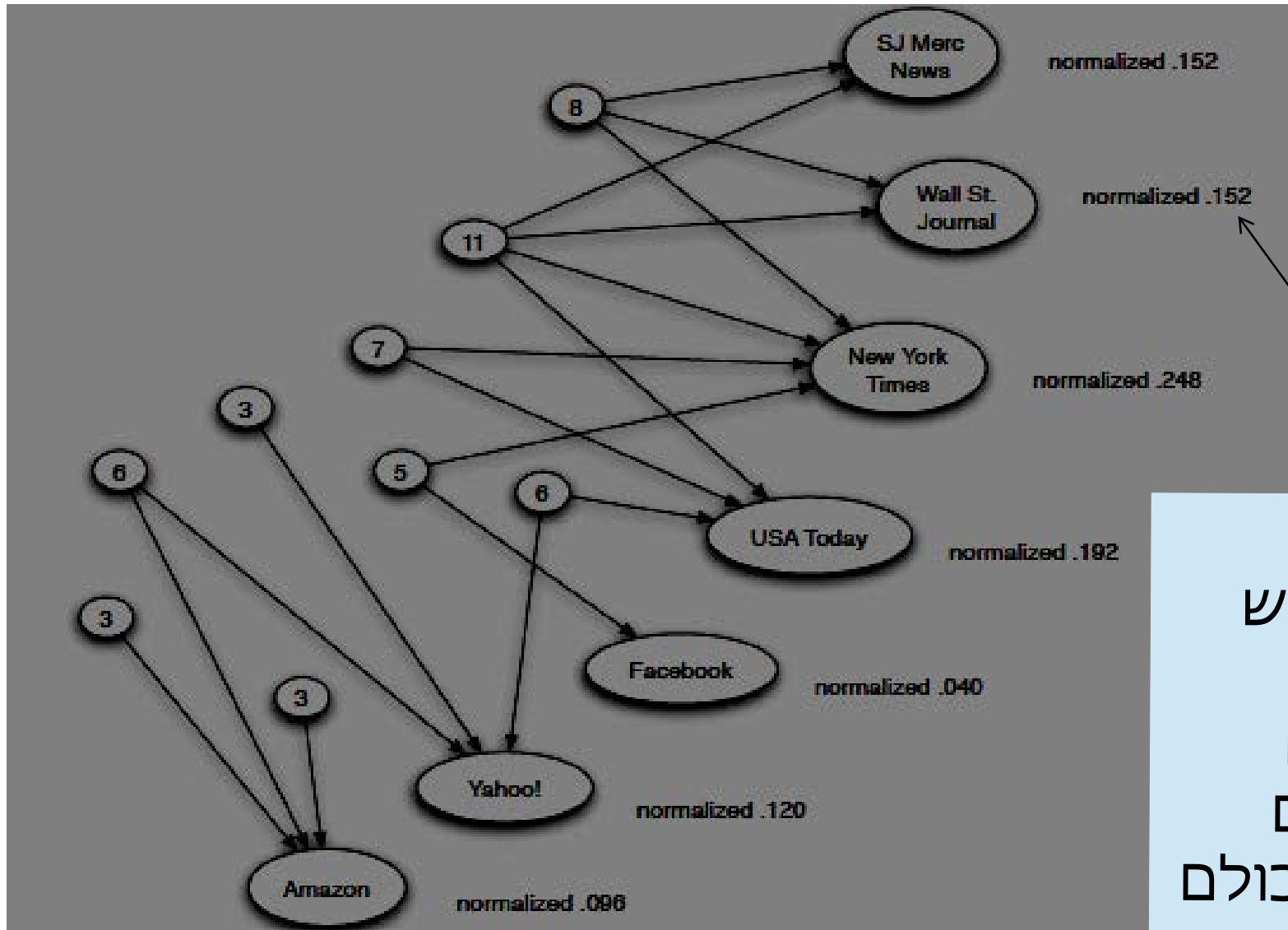
סך הכל קיבלנו

הדפים
המצביעים
לאיבר ברשימה

הדירוג החדש
הוא סכום
המשקלים



לאחר נירמול



הדירוג החדש
הוא סכום
המשקלים
חלקי סכום
הדירוגים של כולם

The Principle Of Repeated Improvement

דפים המדורגים **גבוה** ←

← בעלי **חוש טוב** היכן התוצאות הטובות

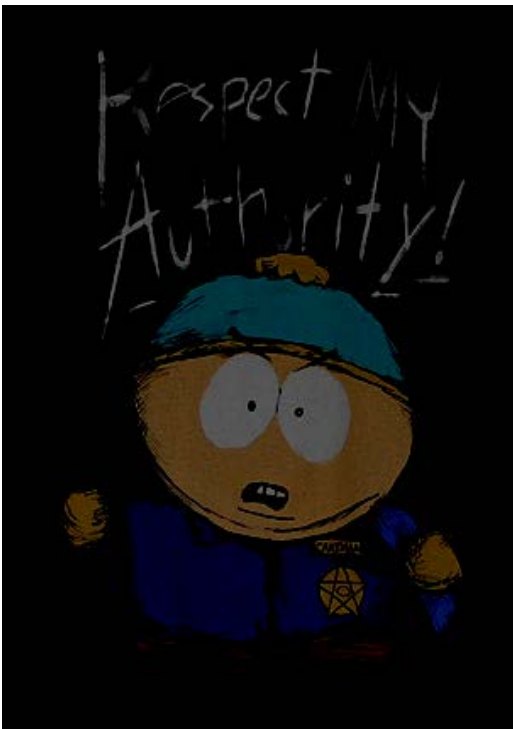
ניתן **משקל גבוה** יותר להצבעות שלהם

Hubs And Authorities

לכל דף ברשת שני תפקידים,

תפקיד כמצביע (Hub)

תפקיד כאיבר ברשימת תוצאות החיפוש (Auth)



Hub הוא חשוב אם הוא **מוביל**

ל- Authorities חשובים

Authority הוא חשוב אם הוא

Hubs חשובים **מובילים** אליו

Hubs And Authorities

כעת לכל דף שנשמנו כ - Pניתן
- ערך Hub פוטנציאלי שנסמן כ -

Hub (P)

- וערך Authority פוטנציאלי שנסמן כ -

.Auth (P)

Hubs And Authorities

התהליך למציאת ערכי Hub ו Authority

הגדרות:

Authority Update Rule-

Hub Update Rule-

Hubs And Authorities

Authority Update Rule (A)

לכל דף נעדכן את ערך ה Authority להיות סכום ערכי ה Hubs של כל הדפים המצביעים עליו

Hubs Update Rule (H)

לכל דף נעדכן את ערך ה Hub להיות סכום ערכי ה Authority של כל הדפים שהוא מצביע עליהם

התהליך למציאת H & A

איתחול כל ערכי ה Hub ו Auth ל- 1

בחירת מספר הצעדים - K

ביצוע K פעמים **H&A Updates** באופן הבא:

Authority Update Rule-

Hub Update Rule-

מה קיבלנו?

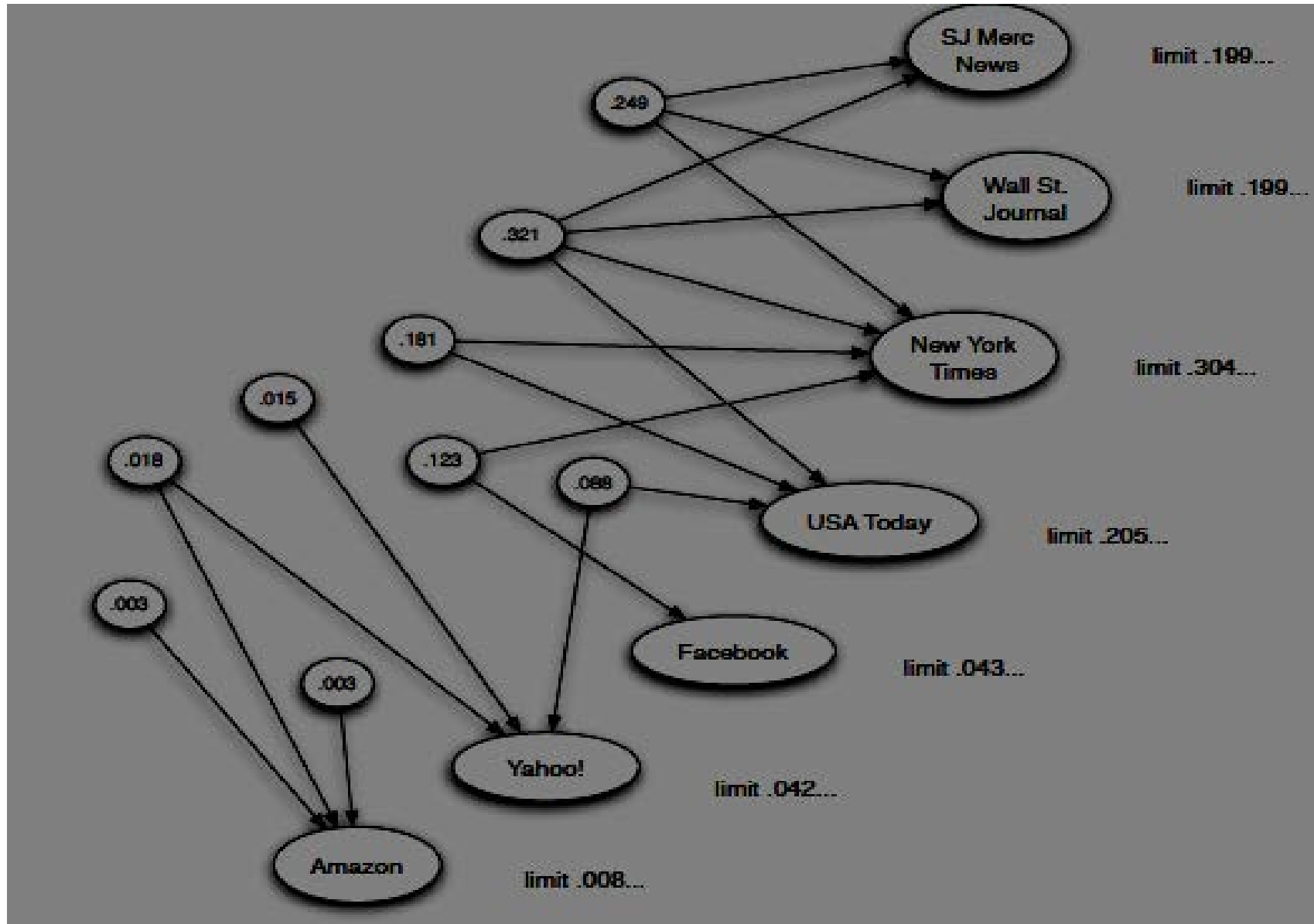
ערך נומרי ל H&A של כל דף

המספרים שקיבלנו מאוד גדולים

אפשר לנרמל על ידי חלוקה:

- לכל ערך Auth נחלק בסכום ערכי ה. Authority
- לכל ערך Hub נחלק בסכום ערכי ה. Hubs

דוגמא לתהליך אחרי נירמול



נקודות חשובות נוספות

• מוביל להתכנסות ההערכים $\rightarrow \infty$.

ההתכנסות מובילה לשיווי משקל ברשת) נגדיר בהמשך)

• התהליך תלוי רק במבנה הרשת

לא נוכיח, אך אפשר להוכיח שאין תלות בערך ההתחלה שניתן ל Auth ו Hub

טכניקה שניה: PageRank

הרעיון:

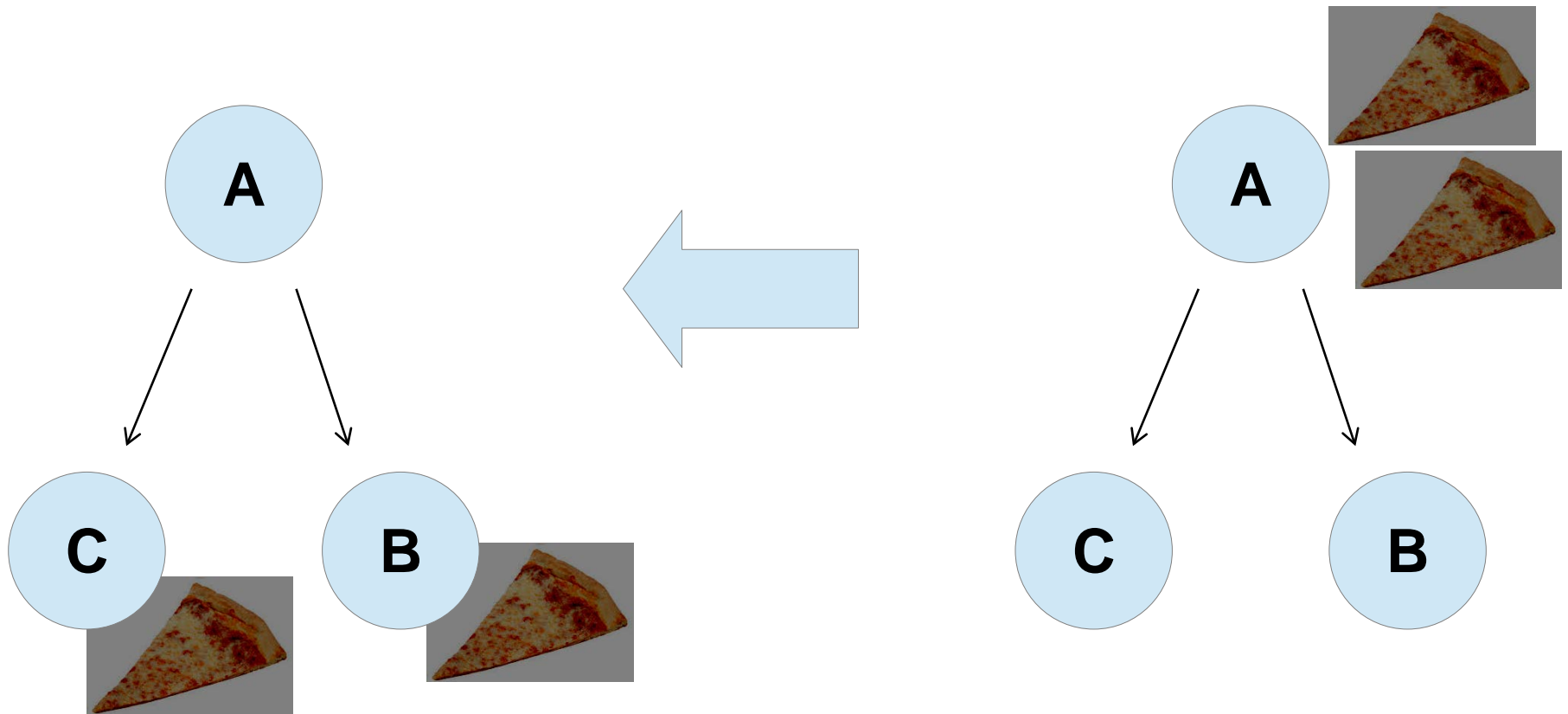
דף הוא **חשוב** אם דפים חשובים **מצביעים** אליו
לדוגמא, מאמרים שמצביעים אחד על השני, בלוגים וכ'

דפים הנקראים חשובים, נעשים **מצביעים חזקים**
יותר

PageRank הוא כמו נוזל ברשת הנאסף בצמתים
החשובים

The Basic Definition Of PageRank

כל צומת מחלק את ה PageRank הנוכחי שלו שווה בשווה בין השכנים שהוא מוביל אליהם



The Basic Definition Of PageRank

התהליך למציאת PageRank ע"י ההגדרה
הבסיסית :

ברשת עם N צמתים, לכל צומת ניתן ערך של

$$.1/N$$

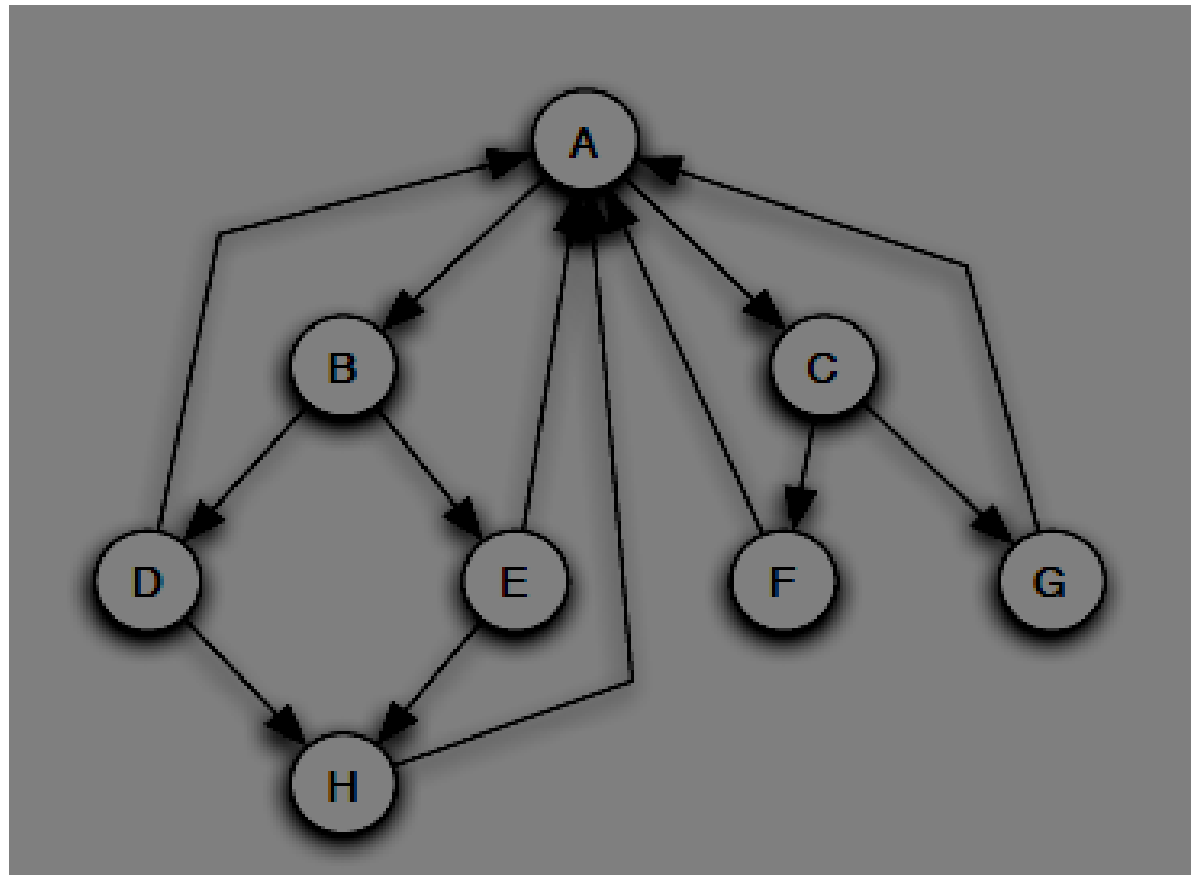
נבחר מספר צעדים K

נבצע אפעמים את

The Basic Definition Update Rule.

דוגמא

נבצע את התהליך לרשת הבאה



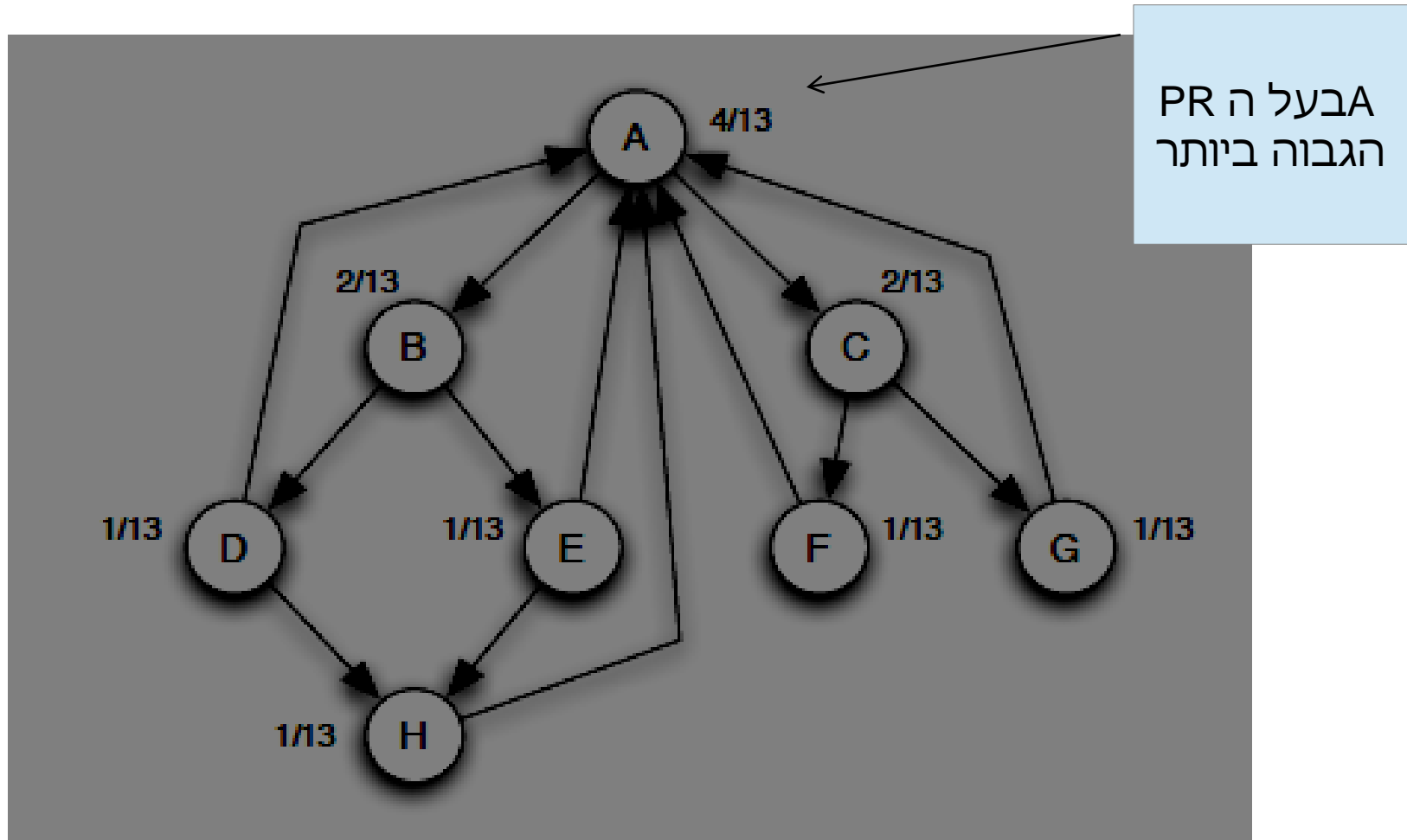
כל צומת
מקבל
 $1/8$

דוגמא

שני השלבים הראשונים בתהליך הבסיסי :

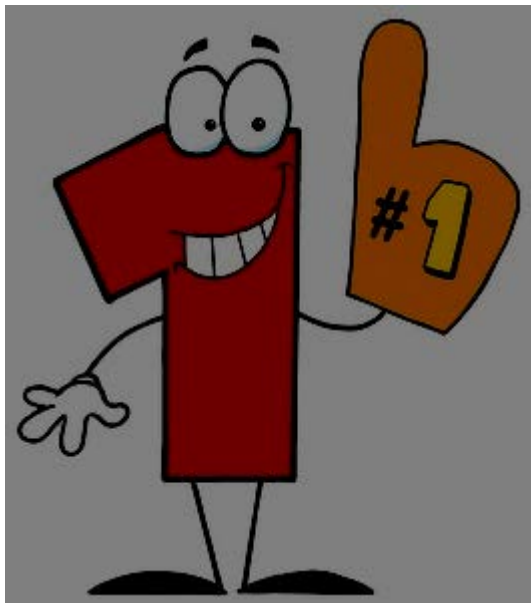
Step	A	B	C	D	E	F	G	H
1	$1/2$	$1/16$	$1/16$	$1/16$	$1/16$	$1/16$	$1/16$	$1/8$
2	$3/16$	$1/4$	$1/4$	$1/32$	$1/32$	$1/32$	$1/32$	$1/16$

לאחר התכנסות התהליך



אין צורך בנירמול

נשים לב שבשיטה זו אין צורך בנירמול כיוון
שתמיד סכום ה Page Ranks הוא 1



Equilibrium Values Of PageRank

הגדרה:

הרשת נמצאת בשיווי משקל אם כאשר מבצעים עדכון לרשת, הערכי ה PageRank - נשמרים

Equilibrium Values Of PageRank

נקודות חשובות:

בתהליך ה Basic Page Rank,

מוביל ל**התכנסות** ההערכים $\rightarrow \infty$.

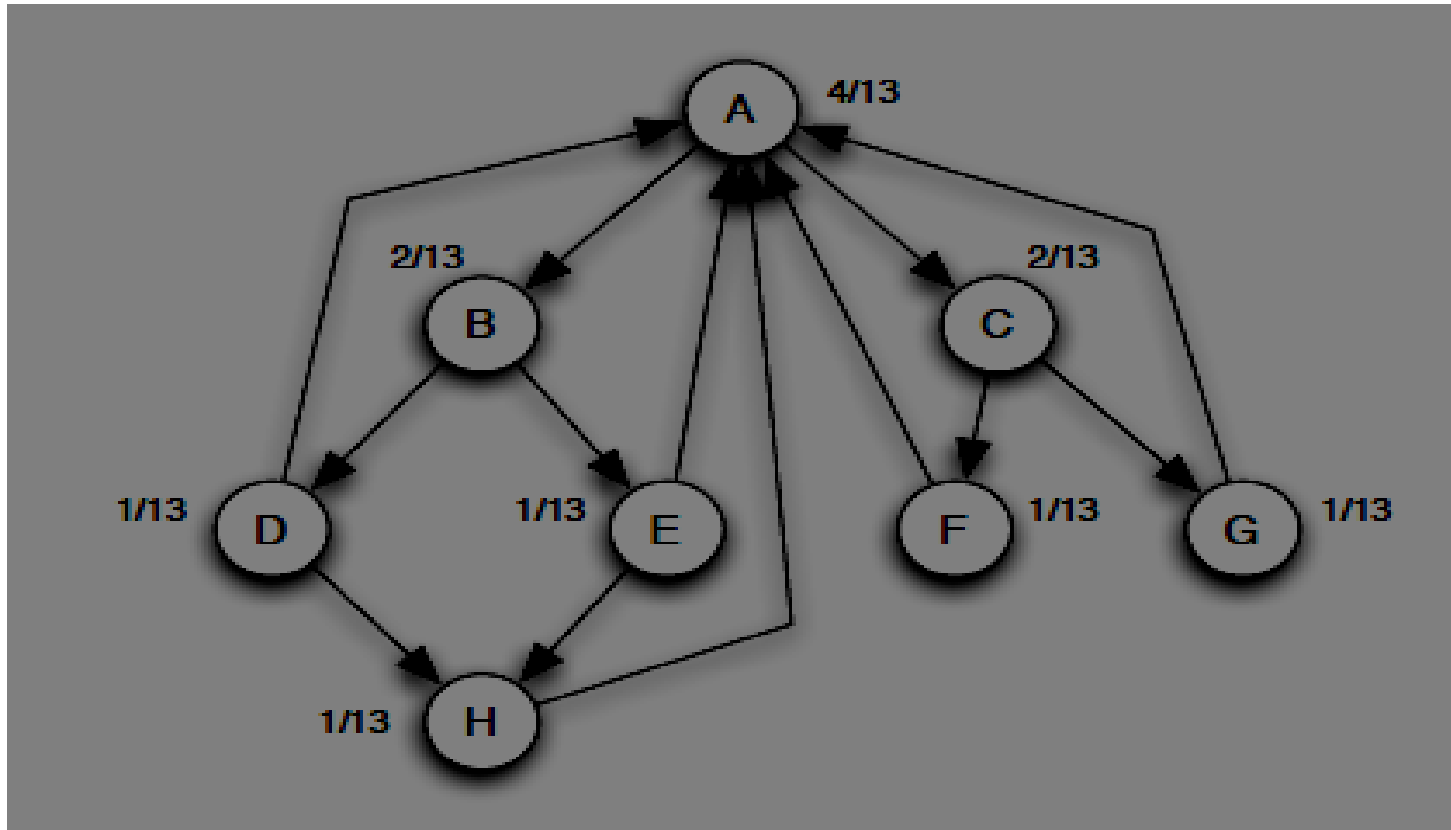
התכנסות מובילה ל**שיווי משקל** ברשת.

איך יודעים שאנו בשיווי משקל?

- סכום המשקלים הוא 1

- נשמר תחת העדכון הבסיסי

בחזרה לדוגמא



הרשת נמצאת בשיווי משקל

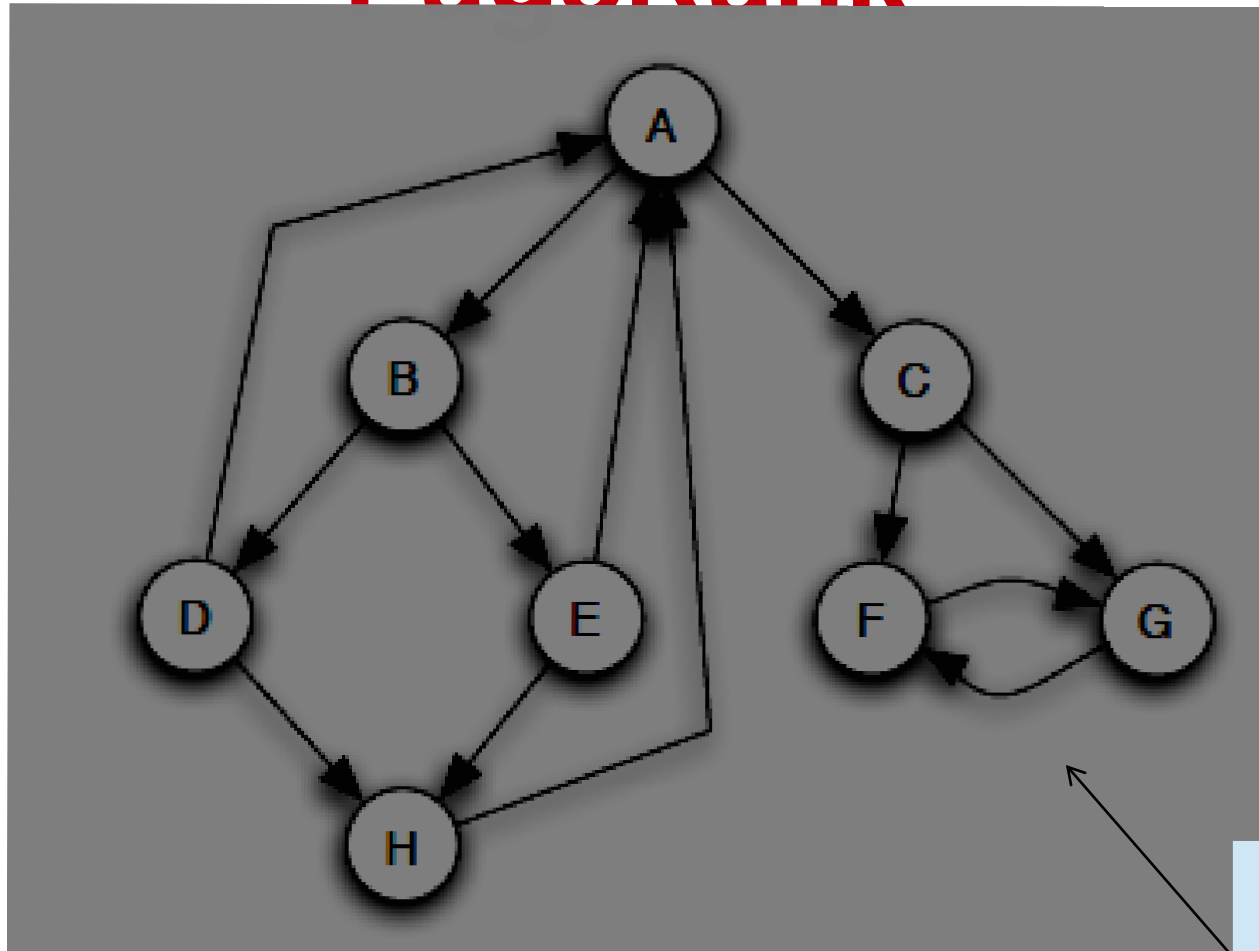
Scaling The Definition Of PageRank

בעיה

צמתים (או רכיבי קשירות) **המכילים** את כל ה
PageRank

← נראה בדוגמא הבאה ע"י שינוי הרשת

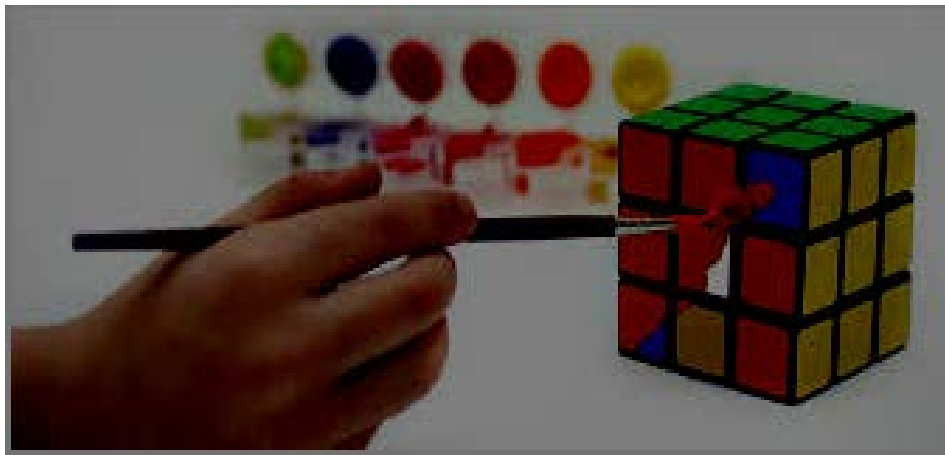
Scaling The Definition Of PageRank



כל ה -
PageRank
יתאסף ב -
F & G

Scaling The Definition Of PageRank

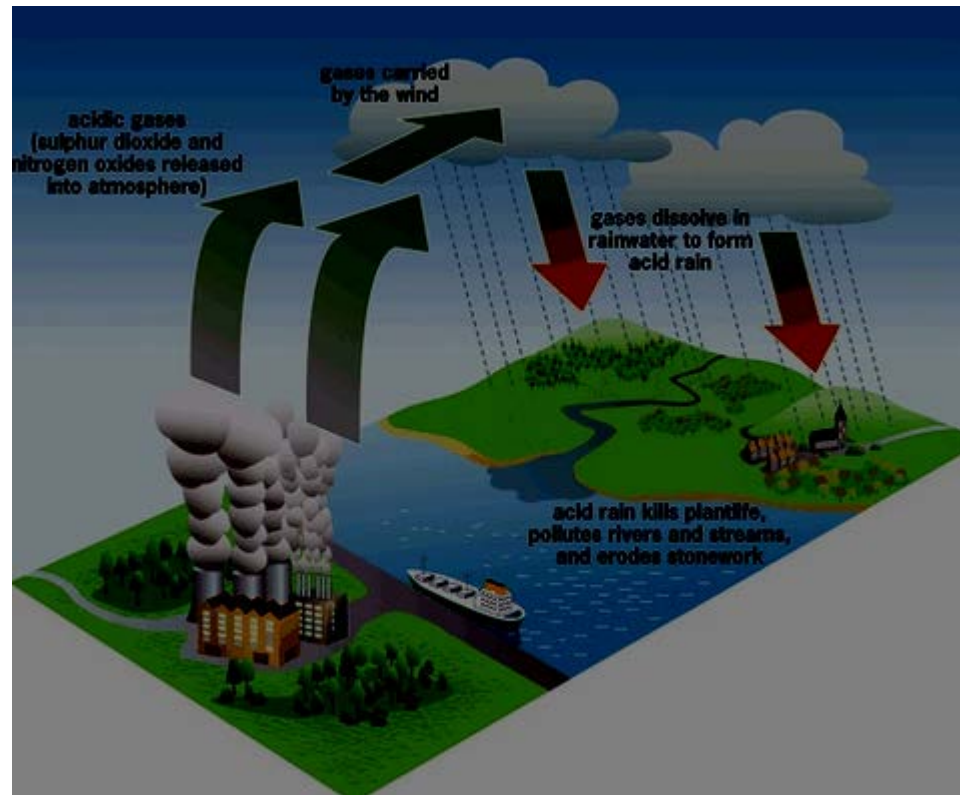
בשיווי משקל של המערכת החדשה,
צמתים F ו- G מקבלים את הערך 0.5 בעוד ששאר
הצמתים מקבלים ערך 0



יש פתרון !

Scaling The Definition Of PageRank

קצת מוטיבציה לפתרון:



הגשם מתפזר גם במקומות הגבוהים שבכדור הארץ

הפתרון לבעיה

הגדרה:

Scaled PageRank Update Rule :

1. נבחר מספר S .

$$0 < S < 1$$

2. נבצע את ה **Basic Update Rule**.

3. נכפיל את כל הערכים שקיבלנו ב $-S$.

הפתרון לבעיה

כיוון שסכום הערכים היה 1 , כעת הוא S
לכן נשארנו עם

$$(1 - S)$$

מה- PageRank

4. כל צומת מקבל תוספת של

$$(1 - S) / N$$

כאשר N הוא מספר הצמתים ברשת

The Limit of the Scale Update

נקודות חשובות:

• - ∞ → מוביל להתכנסות ההערכים
התכנסות מובילה לשיווי משקל ברשת

שיטה זו נעשת בפרקטיקה

• - בדרך כלל בוחרים ערך S שהוא בין 0.8 ל-0.9 (Best Practice)

Random Walks

נשתמש ב Random Walks על מנת להסתכל על
חלוקת ה PageRank בצורה קצת שונה

הגדרה :

Random Walks זהו דרך לטייל ברשת בצורה אקראית,
מתחילים מדף A , ובוחרים באופן אקראי את אחד הלינקים
היוצאים מ A - וכך הלאה

Random Walks - Example

נניח שיש אדם הנכנס לרשת, בוחר מספר צעדים
 K , ומתחיל לגלוש בה בצורה **רנדומלית**,

הוא מתחיל מדף A , **בוחר באופן אקראי** דף מכל
הדפים היוצאים מ A , וממשיך אליו, כך עושה K
פעמים.

Random Walks

טענה:

הסיכוי להיות בדף **X** אחרי **K** צעדים הוא בדיוק
ערך ה **PageRank** של דף **X** אחרי **K** **עדכונים** של
The Basic PageRank Update Rule

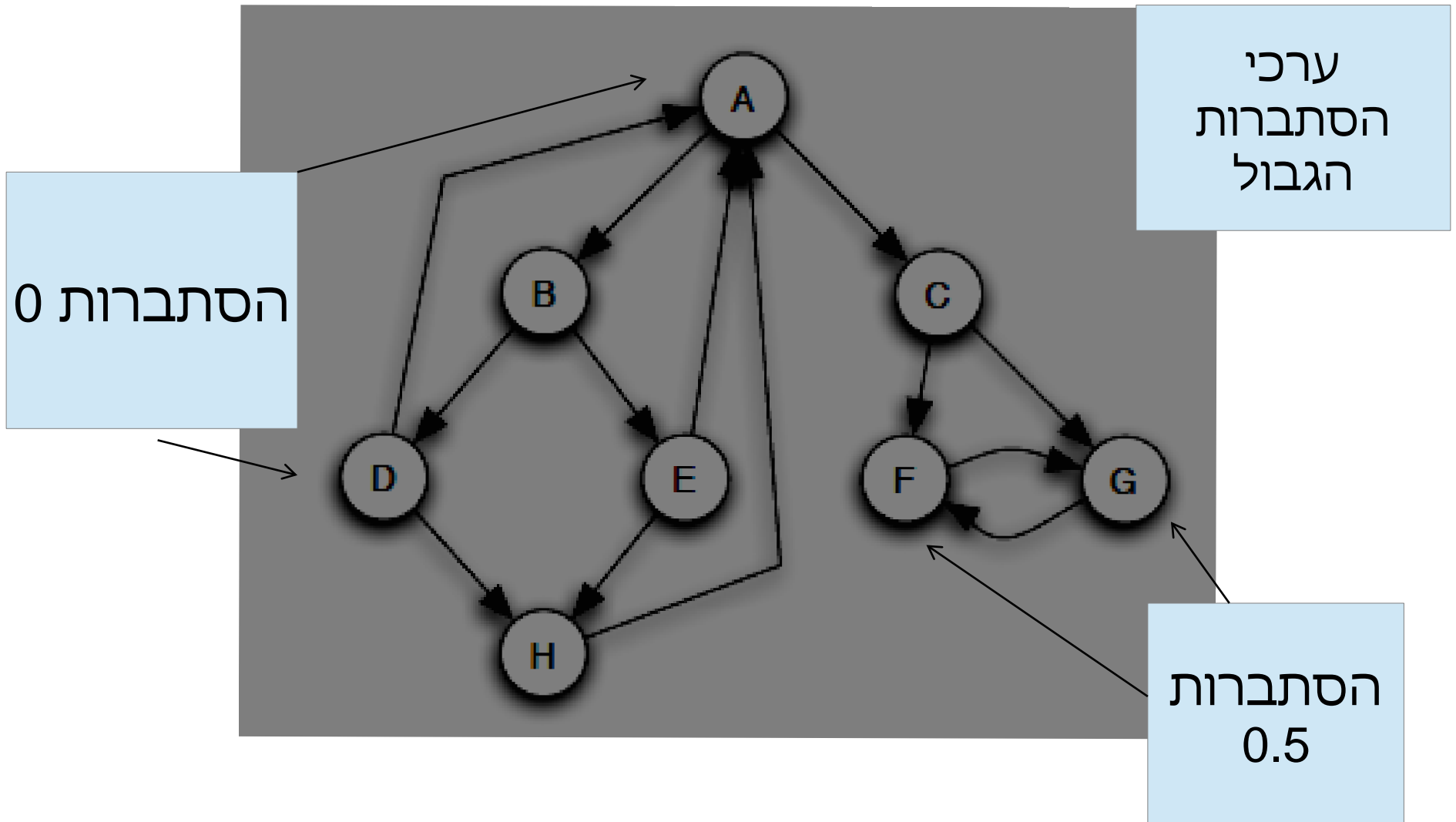
יש כאן שקילות

בהנחה ששתי הנוסחאות לחישוב PageRank
שקולות, Repeated Improvement 1 – Random
Walks

אפשר להסיק:

ה PageRank של דף X הוא **הסתברות הגבול**
שנגיע לדף X על ידי מספר גדול (אינסופי) של
צעדים רנדומליים ברשת.

בחזרה לדוגמא



הערות אחרונות

הנושא מאוד מעניין ואפשר להעמיק בו מאוד

לחברות שונות ישנן שיטות שונות לחישוב

PageRank, אך הן שומרות שיטות אלו בסוד

אפשר לסבך את הרשת ולשאול שאלות קשות יותר

אפשר להשתמש בעוד נתונים מהרשת ולתת

תוצאות אמינות יותר

תודה !

