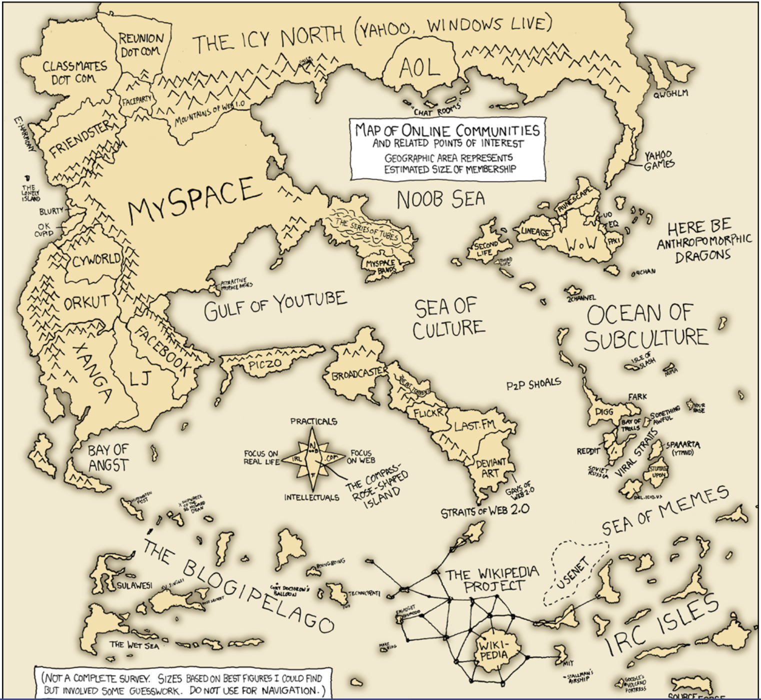
Misleading stars: What cannot be measured on the Internet ?

Yvonne-Anne Pignolet, Stefan Schmid, Gilles Tredan


MAP OF ONLINE COMMUNITIES
AND RELATED POINTS OF INTEREST
GEOGRAPHIC AREA REPRESENTS
ESTIMATED SIZE OF MEMBERSHIP
THE ICY NORTH (YAHOO, WINDOWS LIVE)
REUNION DOT COM
CLASSMATES DOT COM
AOL
"CHAT ROOMS"
QWGHLM
FACEPARTY
MOUNTAINS OF WEB 1.0
FRIENDSTER
E-HARMONY
THE LONELY ISLAND
MYSPACE
YAHOO GAMES
NOOB SEA
BLURTY
OK CUPID
THE SERIES OF TUBES
RUNESCAPE
UO
EQ
LINEAGE
SECOND LIFE
THIRD LIFE
WoW
FFXI
HERE BE ANTHROPOMORPHIC DRAGONS
CYWORLD
MYSPACE BANDS
4CHAN
ATTRACTIVE MYSPACE PAGES
ORKUT
GULF OF YOUTUBE
SEA OF CULTURE
2CHANNEL
OCEAN OF SUBCULTURE
FACEBOOK
XANGA
LJ
PICZO
BROADCASTER
BIT TORRENT
ISLE OF SLASH
P2P SHOALS
FARK
FLICKR
LAST.FM
DIGG
SOMETHING AWFUL
BAY OF TROLLS
PRACTICALS
FOCUS ON REAL LIFE
FOCUS ON WEB
IRL
.COM
INTELLECTUALS
THE COMPASS-ROSE-SHAPED ISLAND
DEVIANT ART
REDDIT
VIRAL STRAITS
SPAAARTA (YTMND)
SOVIET RUSSIA
STUMBLE UPON
BAY OF ANGST
GAYS OF WEB 2.0
DEL.ICIO.US
STRAITS OF WEB 2.0
SEA OF MEMES
THE BLOGIPELAGO
BOINGBOING
SULAWESI
CORY DOCTOROW'S BALLOON
TECHNORATI
THE WIKIPEDIA PROJECT
USENET
THE WET SEA
WIKI-PEDIA
IRC ISLES
MIT
STALLMAN'S AIRSHIP
GOOGLE'S VOLCANO FORTRESS
SOURCEFORGE
(NOT A COMPLETE SURVEY. SIZES BASED ON BEST FIGURES I COULD FIND BUT INVOLVED SOME GUESSWORK. DO NOT USE FOR NAVIGATION.)

How accurate are network maps ?

Why ?

- To develop/adapt protocols to Internet *PaDIS, RMTP*
- To understand the impact of uncertainty: *networks metrology ?*

How ?

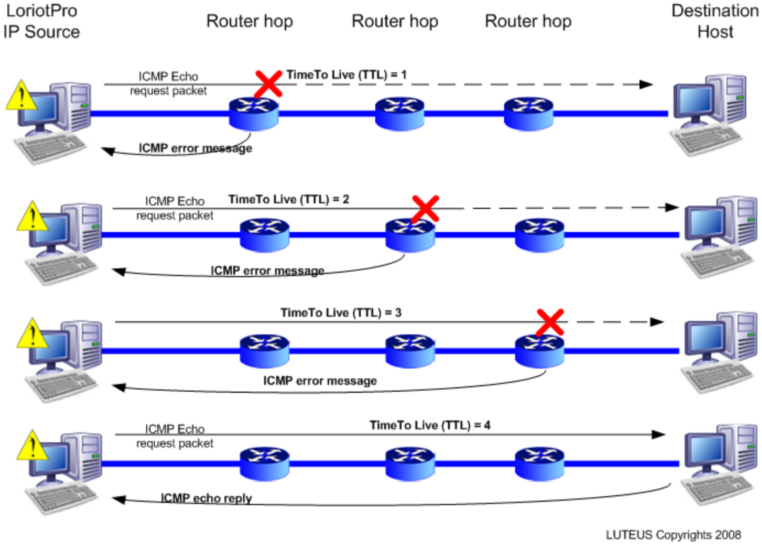
- multicast [*Marchetta* et al. , *JSAC*′11]
- network tomography
- Traceroute

Takeaway

- Compare internet topologies instead of counting them
- Local properties suffer
- Global properties suffer less

Traceroute principle

Idea= send « buggy » TTL packets, collect error messages



Traceroute Limitations

- Sampling bias: Impact of sources location
- Aliasing: How to map IPs to the equipement
- Load-Balancing: Traceroute assumes all packets follow the same path...
- Stars: Disabled/Filtered ICMP messages

$\longrightarrow$ Trace:

$t_1 = a, \star_1, b$

$t_2 = b, \star_2, i_1, i_2, i_3, \star_3, c$

Traceroute Limitations

- Sampling bias: Impact of sources location
- Aliasing: How to map IPs to the equipement
- Load-Balancing: Traceroute assumes all packets follow the same path...
- Stars: Disabled/Filtered ICMP messages

$\longrightarrow$ Trace:

$t_1 = a, \star_1, b$

$t_2 = b, \star_2, i_1, i_2, i_3, \star_3, c$

Traceroute Limitations

- Sampling bias: Impact of sources location
- Aliasing: How to map IPs to the equipement
- Load-Balancing: Traceroute assumes all packets follow the same path...
- Stars: Disabled/Filtered ICMP messages

Trace:

$t_1 = a, \star_1, b$

$t_2 = b, \star_2, i_1, i_2, i_3, \star_3, c$ $\longleftarrow$

?

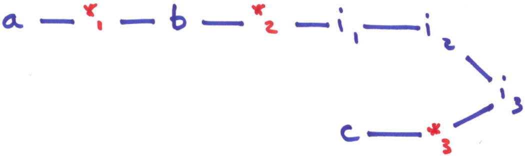
Traceroute Limitations

- Sampling bias: Impact of sources location
- Aliasing: How to map IPs to the equipement
- Load-Balancing: Traceroute assumes all packets follow the same path...
- Stars: Disabled/Filtered ICMP messages

Trace:

$t_1 = a, \star_1, b$

$t_2 = b, \star_2, i_1, i_2, i_3, \star_3, c$ $\longleftarrow$

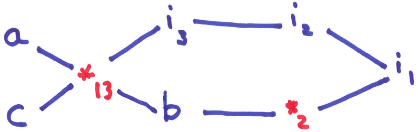


- Sampling bias: Impact of sources location
- Aliasing: How to map IPs to the equipement
- Load-Balancing: Traceroute assumes all packets follow the same path...
- Stars: Disabled/Filtered ICMP messages

Trace:

$t_1 = a, \star_1, b$

$t_2 = b, \star_2, i_1, i_2, i_3, \star_3, c$ $\longleftarrow$



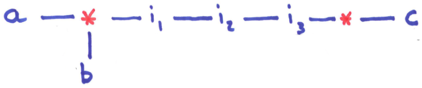
Traceroute Limitations

- Sampling bias: Impact of sources location
- Aliasing: How to map IPs to the equipement
- Load-Balancing: Traceroute assumes all packets follow the same path...
- Stars: Disabled/Filtered ICMP messages

Trace:

$t_1 = a, \star_1, b$

$t_2 = b, \star_2, i_1, i_2, i_3, \star_3, c$ $\longleftarrow$



Related Works

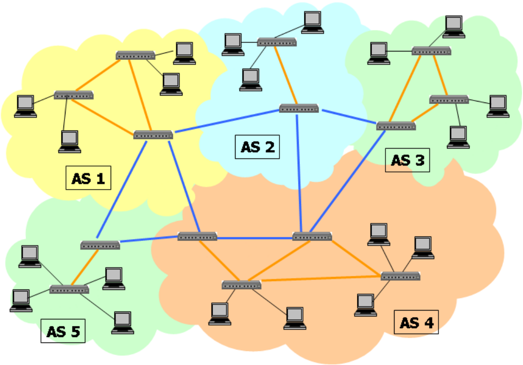
Works initiated by Acharya and Gouda [SSS09, $ICDCN$10, $ICDCN$11]

- Models to capture *irregular* nodes
- Irregular nodes $\rightarrow$ anonymous nodes
- Central concept of *minimal* topologies
- Counting the number of generable topologies

Our approach: Lots of generable topologies is not a problem if they are all similar

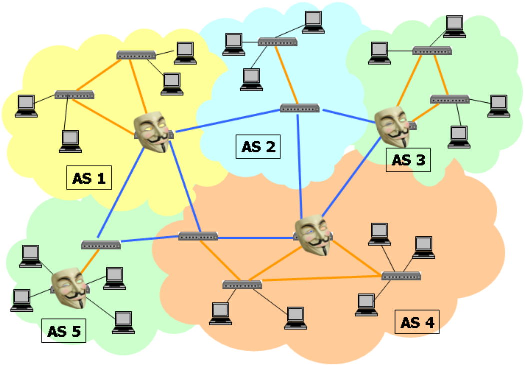
Model

- $G_0(V_0, E_0)$: static undirected graph = **Target Topology**.
- $v \in V_0$ is
  - either named: always answers with its only name (no aliasing)
  - either anonymous: always answers $\star$.
- Not necessarily minimal!
- $d_{G_0}(u, v)$ is the shortest path distance.



Model

- $G_0(V_0, E_0)$: static undirected graph = **Target Topology**.
- $v \in V_0$ is
  - either named: always answers with its only name (no aliasing)
  - either anonymous: always answers $\star$.
- Not necessarily minimal!
- $d_{G_0}(u, v)$ is the shortest path distance.

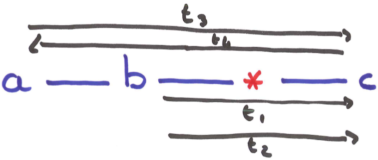


We don't know G_0, but we know a set of traces $\mathcal{T}$ of G_0.

- Anonymous nodes appear as stars ($\star$) in $\mathcal{T}$
- Each star has a unique number ($\star_i, i = 1..s$) in $\mathcal{T}$
- $d_T(u, v)$ is the number of symbols in $T \in \mathcal{T}$ between u and v.
- No Assumption on the number of traces, on path uniqueness nor symmetry.

$\leftrightarrow$

$$t_1 = a, \star_1, b$$
$$t_2 = b, \star_2, i_1, i_2, i_3, \star_3, c$$



$\leftrightarrow$

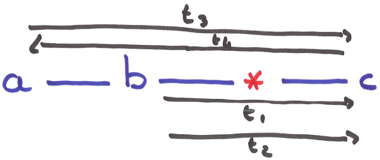
$$t_1 = b, \star_1, c$$
$$t_2 = b, \star_2, c$$
$$t_3 = a, b, \star_3, c$$
$$t_4 = c, \star_4, b, a$$

Rules

Complete cover: Each edge of G_O appears at least once in some trace to $\mathcal{T}$

Reality sampling: For every trace $T \in \mathcal{T}$, if the distance between two symbols $\sigma_1, \sigma_2 \in T$ is $d_T(\sigma_1, \sigma_2) = k$, then there exists a path (i.e., a walk without cycles) of length k connecting two (named or anonymous) nodes σ_1 and σ_2 in G_0.

No assumption on coverage



~~a, b, c~~

~~a, c, b~~

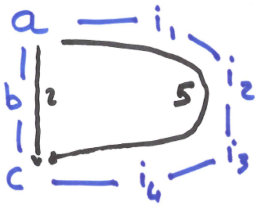
~~$a, \star_1, c$~~

~~a, d, c~~

Rules -2

α-**(Routing) Consistency**: There exists an $\alpha \in (0,1]$ such that, for every trace $T \in \mathcal{T}$, if $d_T(\sigma_1, \sigma_2) = k$ for two entries σ_1, σ_2 in trace T, then the shortest path connecting the two (named or anonymous) nodes corresponding to σ_1 and σ_2 in G_0 has distance at least $\lceil \alpha k \rceil$.

Note: $\alpha > 0 \Leftrightarrow$ loop-less routing



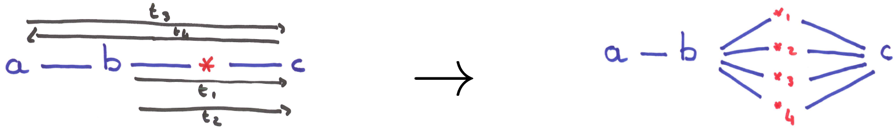
a, b, c

$a, i_1, i_3, i_4, c \Rightarrow \alpha \leq \frac{2}{5}$

A topology G is α-consistently **inferrable** from $\mathcal{T}$ if it respects the 3 previous rules.

Let $\mathcal{G}_{\mathcal{T}} = \{G$, s.t. G is inferrable from $\mathcal{T}\}$ We study the properties of the set $\mathcal{G}_{\mathcal{T}}$ of inferrable topologies.

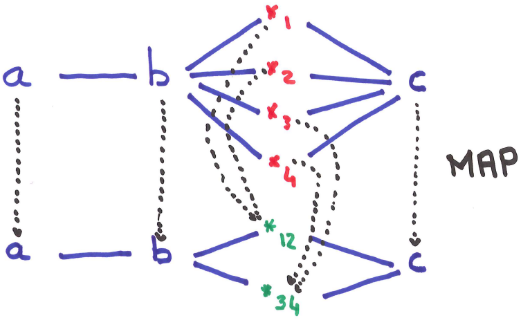
We define the canonic graph G_c as the straightforward graph that treats each star as unique.



How to infer topologies ?

1 inferrable topology= 1 mapping of stars to anonymous routers.

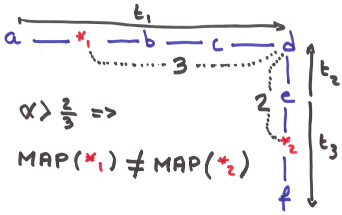
Let Map be such function.



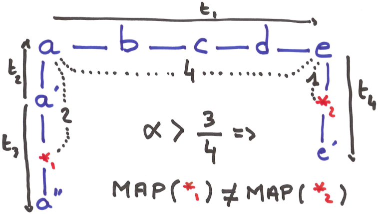
How to infer topologies -2

- G_c is inferrable. Map$=$ Id.

(i) if $\star_1 \in T_1$ and $\star_2 \in T_2$, and $\lceil \alpha \cdot d_{T_1}(\star_1, u) \rceil > d_C(u, \star_2) \Rightarrow$ Map$(\star_1) \neq$ Map$(\star_2)$.



(ii) $\star_1 \in T_1$ $\star_2 \in T_2$, and $\exists T$ s.t. $\lceil \alpha \cdot d_T(u, v) \rceil > d_C(u, \star_1) + d_C(v, \star_2) \Rightarrow$ Map$(\star_1) \neq$ Map$(\star_2)$.



Algorithm constructive part

We construct the *Star Graph* $G_*(V_*, E_*)$:

- Vertices=stars in the trace
- Edges=if stars cannot be merged: $(\star_1, \star_2) \in E_* \Leftrightarrow \text{Map}(\star_1) \neq \text{Map}(\star_2)$.

1 proper coloring of G_* $\leftrightarrow$ 1 Map function

- minimal coloring $\rightarrow$ minimal topology
- 'maximal' coloring $\rightarrow$ G_c

$$\sum_{k=\gamma(G_*)}^{|V_*|} P(G_*, k)/k! \geq |\mathcal{G}_{\mathcal{T}}|,$$

$\gamma(G_*)$ = chromatic number of G_*
$P(G_*, k)$ = *chromatic polynomial* of G_*.

Star Graph /2

→

Trace:

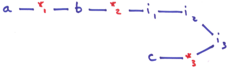
$t_1 = a, \star_1, b$

$t_2 = b, \star_2, i_1, i_2, i_3, \star_3, c$

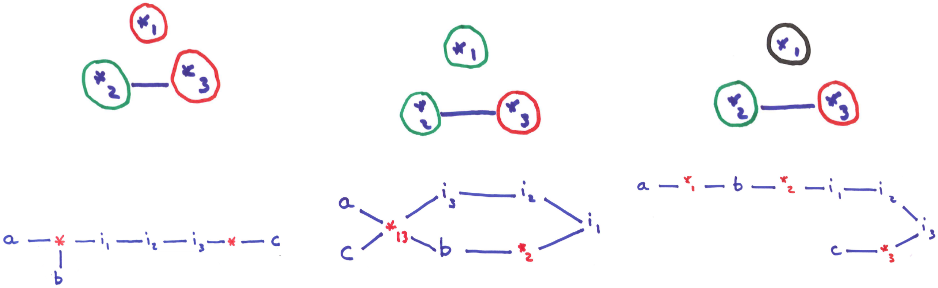
$\alpha = 0.5$

→

G_c :



G_* Colorings



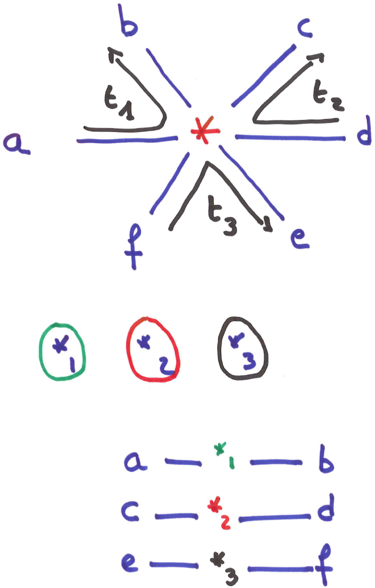
Results are bad !

- **Connected components**
  No assumption on « coverage » ⇒ Stars disconnect the graph !

$$|cc(G_1)/cc(G_2)| \leq \frac{n}{2}$$

- **Stretch**
  Even if we only consider connected topologies.

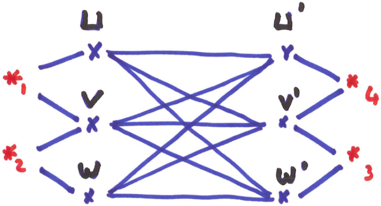
$$|stretch(G)| \leq \frac{n+s-1}{2}$$



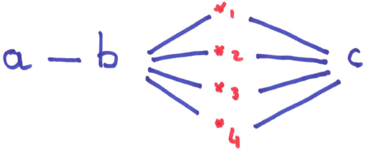
Looking at properties/ Local

It's **worse !**

- **Triangles**
  Bipartite complete graph worst case
  $|C_3(G_1)/C_3(G_2)| \leq \infty$



- **Degree**
  Worst case is on anonymous nodes $|\mathtt{DEG}(G_1) - \mathtt{DEG}(G_2)| \leq 2(s - \gamma(G_\star))$



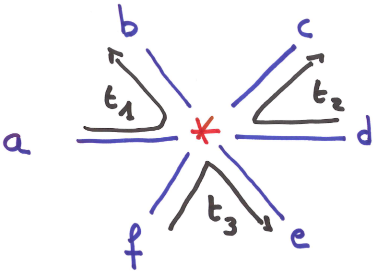
Best case: Fully explored topologies

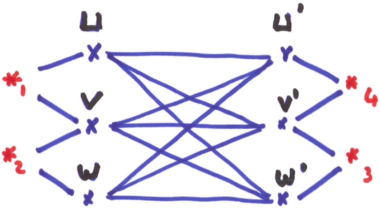
Strong assumptions:

- $\alpha = 1$: shortest path routing
- $\forall u, v \in V_0, \exists T \in \mathcal{T}$ such that $u \in T \vee v \in T$
- We don't see any stronger case

Results:

- Global properties conserved
- Local properties: still bad !





Overall Results

Absolute difference: $G_1 - G_2$

Property	Arbitrary	Fully Explored ($\alpha = 1$)
# of nodes	$\leq s - \gamma(G_*)$	$\leq s - \gamma(G_*)$
# of links	$\leq 2(s - \gamma(G_*))$	$\leq 2(s - \gamma(G_*))$
# of CC	$\leq n/2$	$= 0$
Diameter	$\leq (s-1)/s \cdot (N-1)$	$s/2$ (¶)
Max. Deg.	$\leq 2(s - \gamma(G_*))$	$\leq 2(s - \gamma(G_*))$
Triangles	$\leq 2s(s-1)$	$\leq 2s(s-1)/2$

Relative difference: G_1/G_2

# of nodes	$\leq (n+s)/(n+\gamma(G_*))$	$\leq (n+s)/(n+\gamma(G_*))$
# of links	$\leq (\nu+2s)/(\nu+2)$	$\leq (\nu+2s)/(\nu+2)$
# of CC	$\leq n/2$	$= 1$
Stretch	$\leq (N-1)/2$	$= 1$
Diameter	$\leq s$	2
Max. Deg.	$\leq s - \gamma(G_*) + 1$	$\leq s - \gamma(G_*) + 1$
Triangles	∞	∞

Conclusion

- Constructive proofs
- Complex algorithm
- Don't count, compare!
- Huge dissimilarities
- Worst case approach

A practical part:

- compare with reality
- develop property estimation algorithms

Thanks!