Master Course
Computer Networks
IN2097

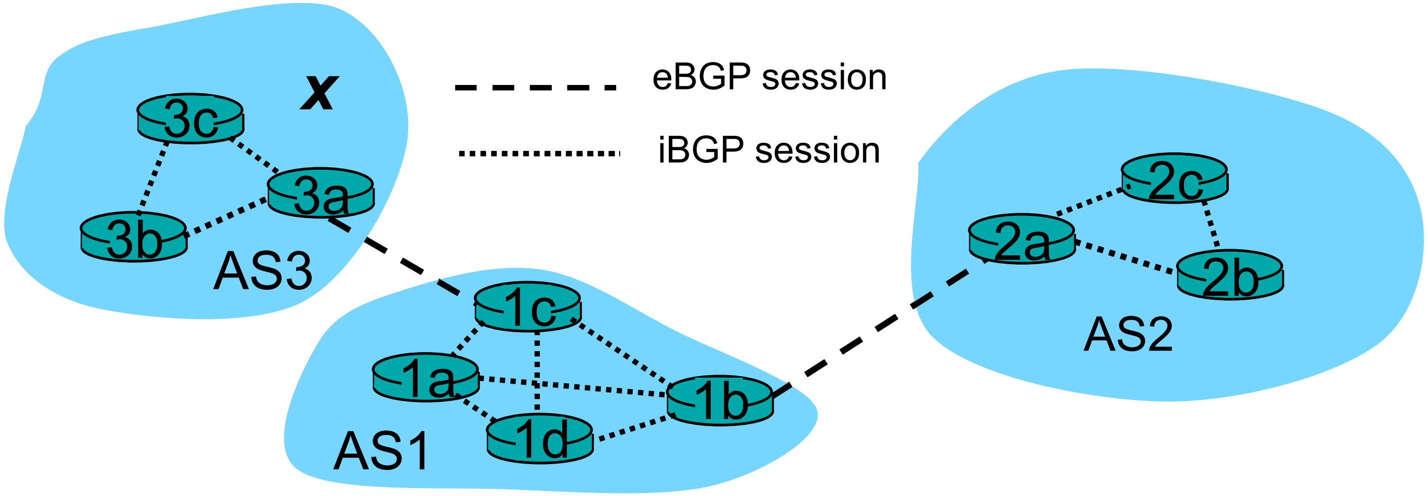
Prof. Dr.-Ing. Georg Carle
Christian Grothoff, Ph.D.
Stephan Günther

Chair for Network Architectures and Services
Department of Computer Science
Technische Universität München
http://www.net.in.tum.de

Routing

eBGP and iBGP

- External BGP: between routers in *different* ASes
- Internal BGP: between routers in *same* AS
  - full IBGP mesh, or route reflectors, or confederations
- No different protocols—just slightly different configurations

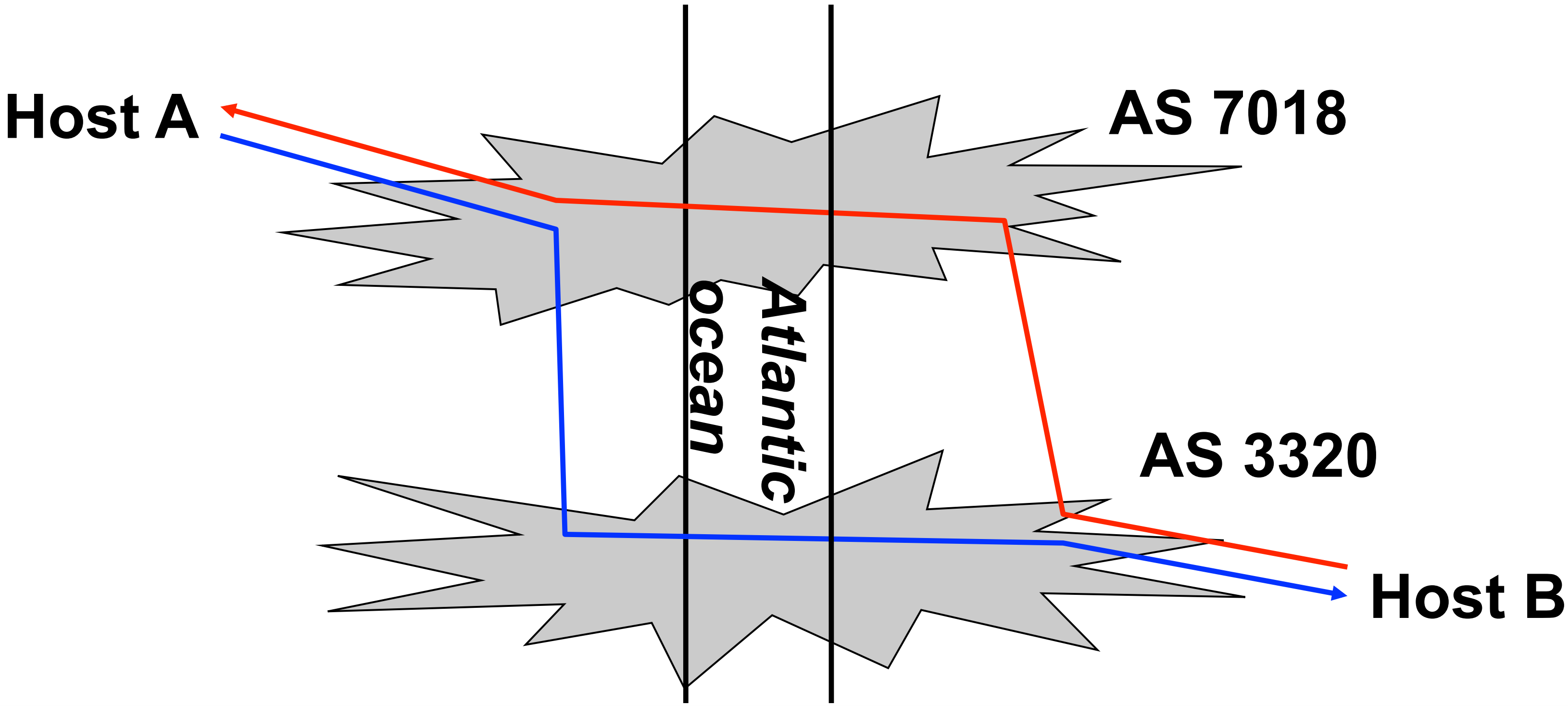


BGP Updates

- Update (Announcement) message consists of
  - Destination (IP prefix)
  - AS Path (=Path vector)
  - Next hop (=IP address of our router connecting to other AS)
- further attributes:
  - Local Preference
  - Origin
  - MED: Multi-Exit Discriminators
  - Community
- More than just path vector protocol
  - In absence of policies, BGP operates with route costs equal to AS_PATH length

Business and Hot-potato routing

- Multiple transit points ⇒ asymmetrical routing
  - Asymmetrical paths very common on the Internet

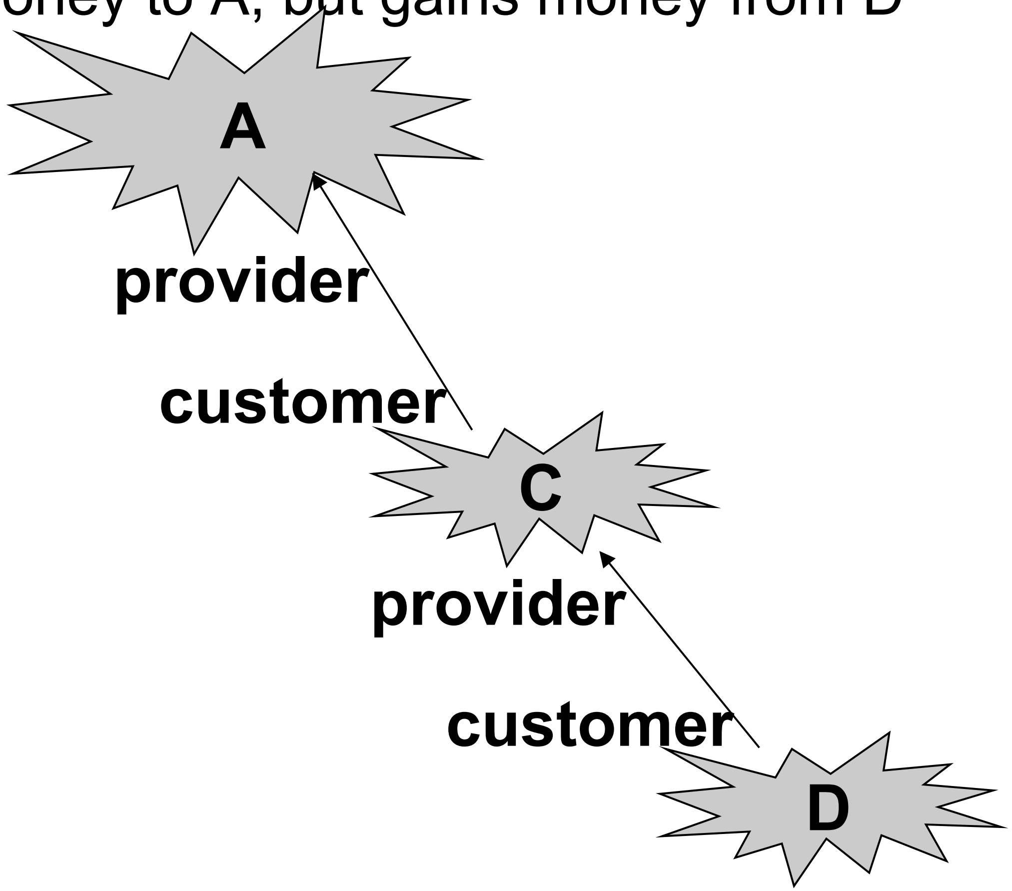


Business and policy routing (1)

- Basic principle #1 (Routing)
  - Prefer routes that incur financial gain
  - …routes via a customer…
  - …are better than routes via a peer, which…
  - …are better than routes via a provider.
- Basic principle #2 (Route announcement)
  - Announce routes that incur financial gain if others use them
    - Others = customers
  - Announce routes that reduce costs if others use them
    - Others = peers
  - Do not announce routes that incur financial loss (…as long as alternative paths exist)

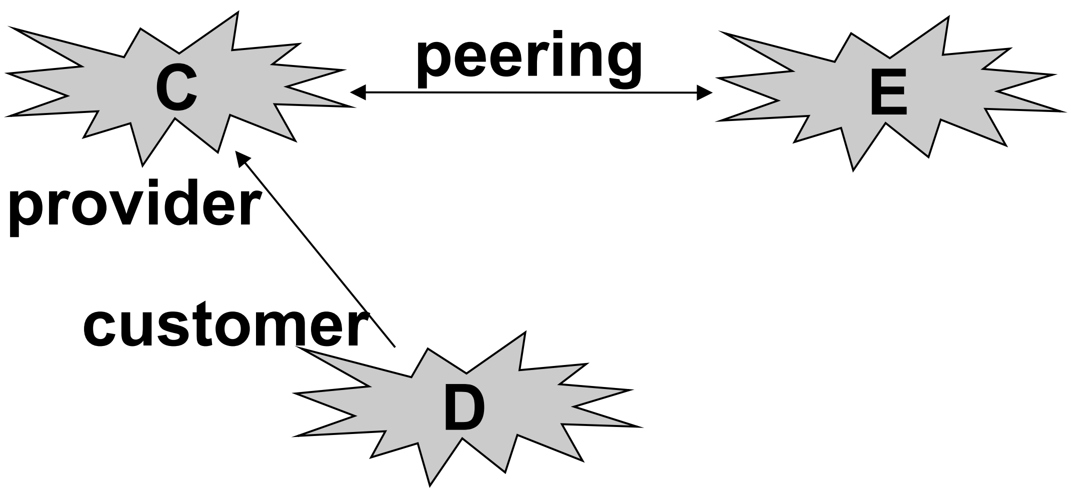
Business and policy routing (6)

- What should C announce here?
  - C tells A about its own prefixes
  - C tells A about its route to D's prefixes: loses money to A, but gains money from D



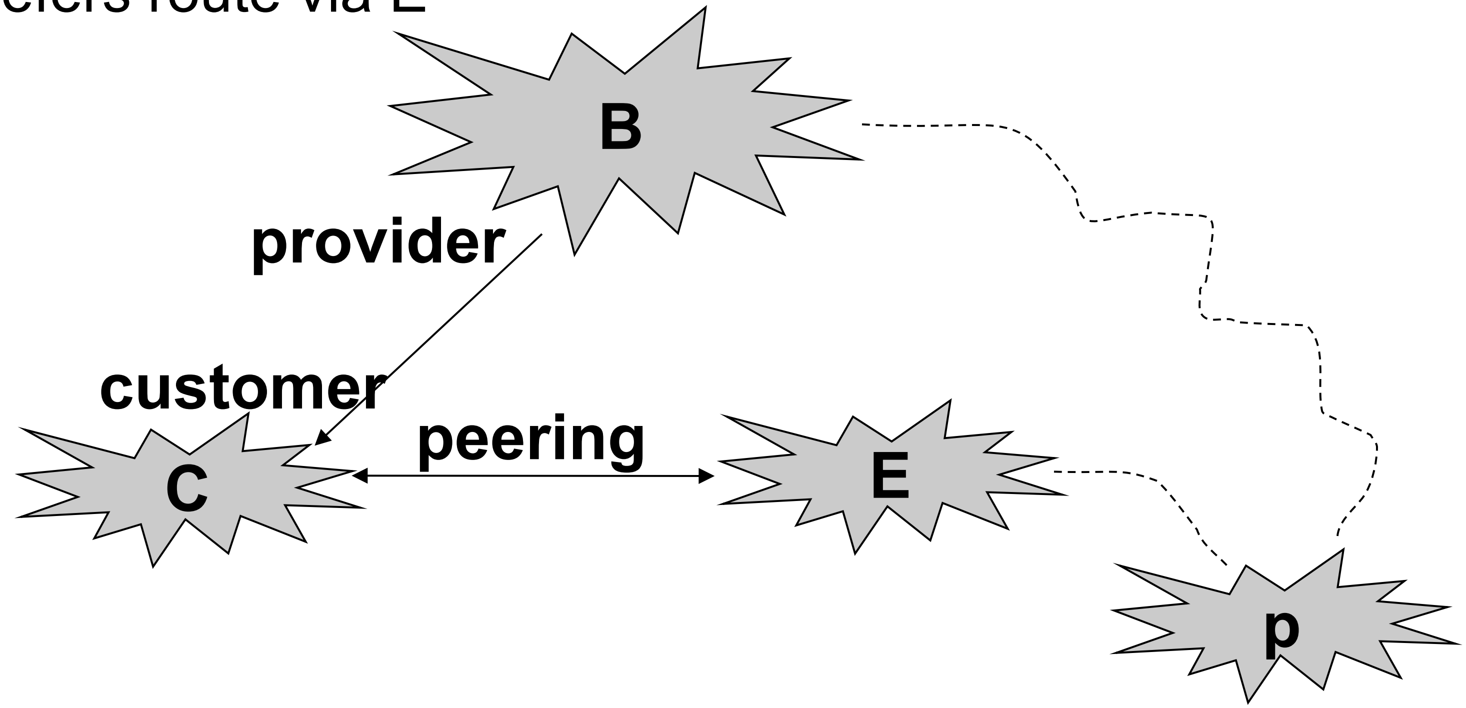
Business and policy routing (7)

- What should C announce here?
  - C tells peering partner E about its own prefixes and route to D:
    no cost on link to E, but gains money from D



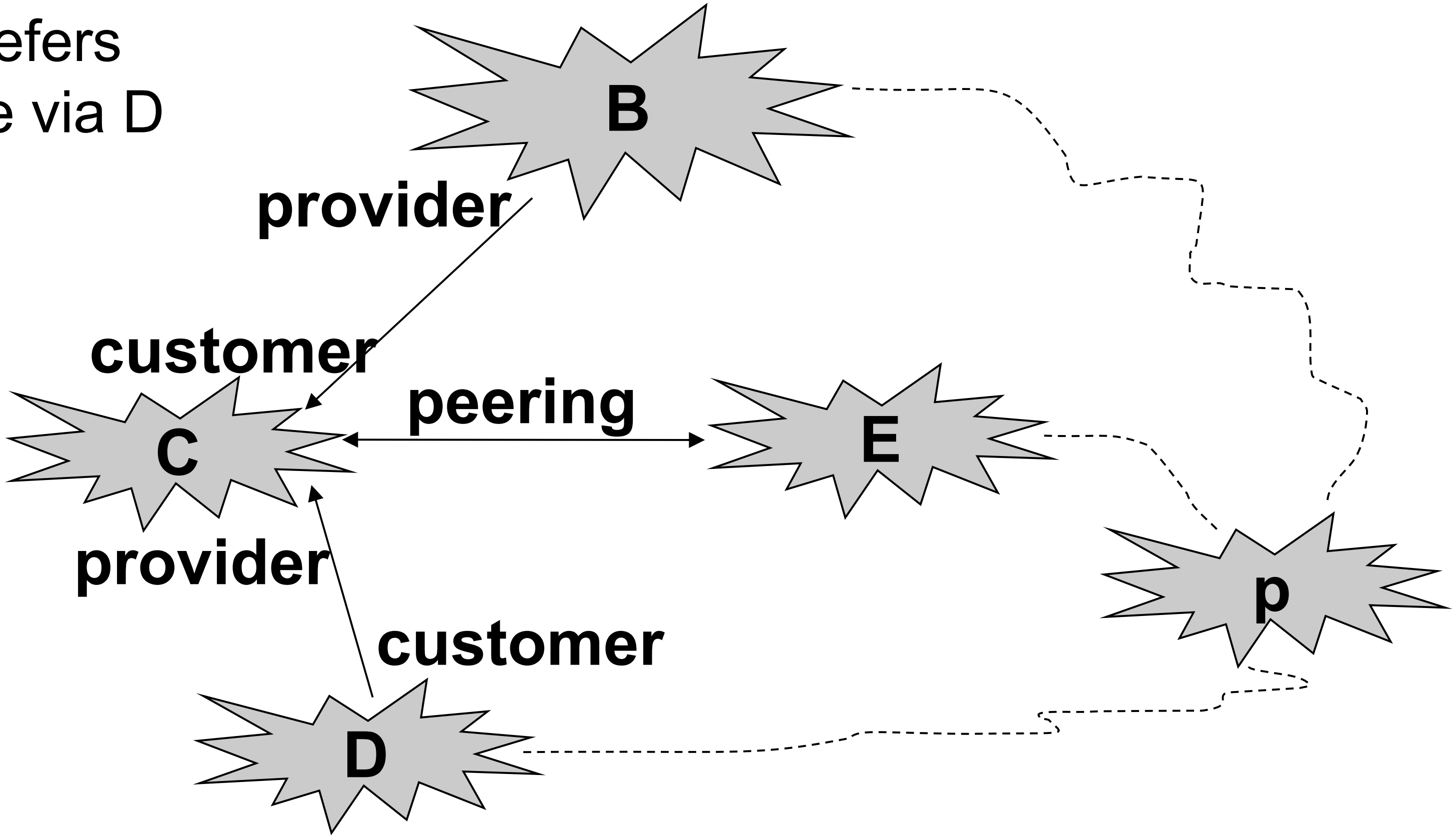
Business and policy routing (8a)

- Which route should C select?
  - B tells C about route to prefix p (lose money)
  - E tells C about route to prefix p (± 0)
  - C prefers route via E



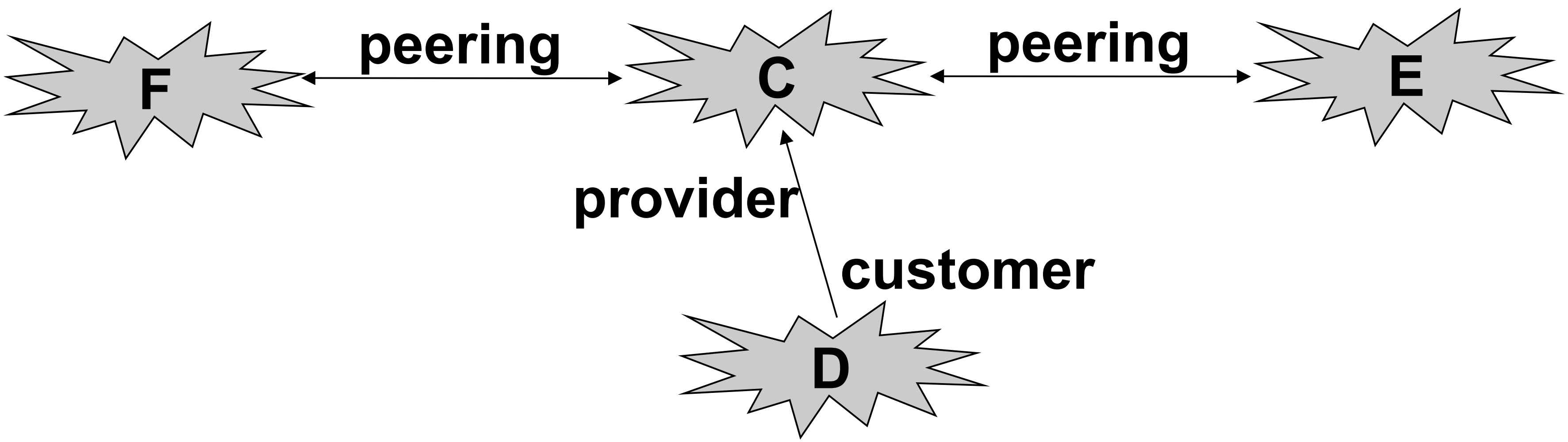
Business and policy routing (8b)

- Which route should C select?
  - B tells C about route to prefix p (lose money)
  - E tells C about route to prefix p (± 0)
  - D tells C about route to prefix p (gain money)
  - C prefers route via D



Business and policy routing (9)

- What should C announce here?
  - C announces to F and E: its own prefixes and D’s routes
  - C does *not* announce to E: routes going via F
    - Otherwise: E could send traffic towards F but wouldn’t pay anything, F wouldn’t pay either, and C’s network gets loaded with additional traffic
  - C does *not* announce to F: routes going via E
    - Same reason

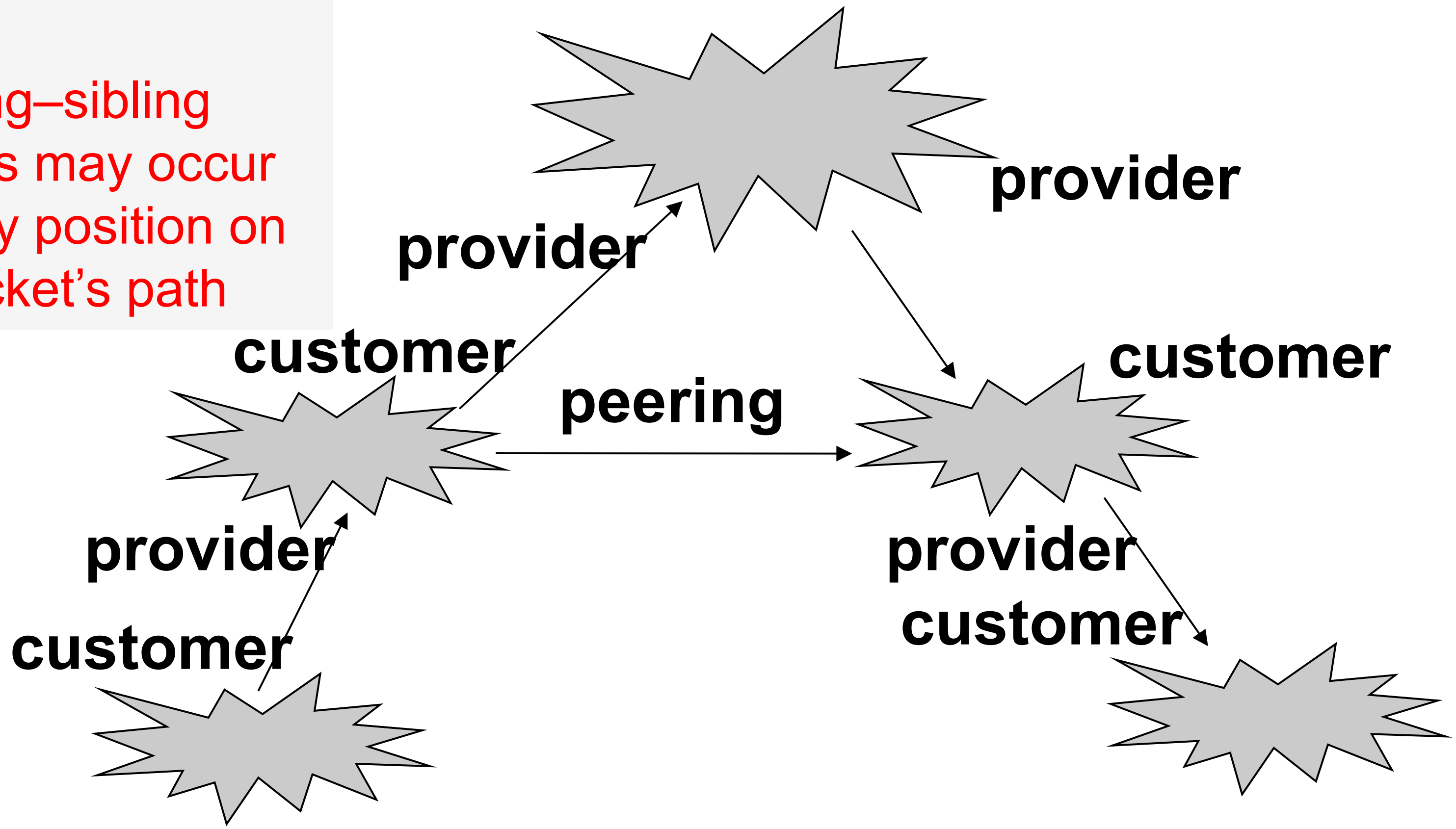


Policy routing: Valley-free routing (idealised!)

Results: Packets always travel…

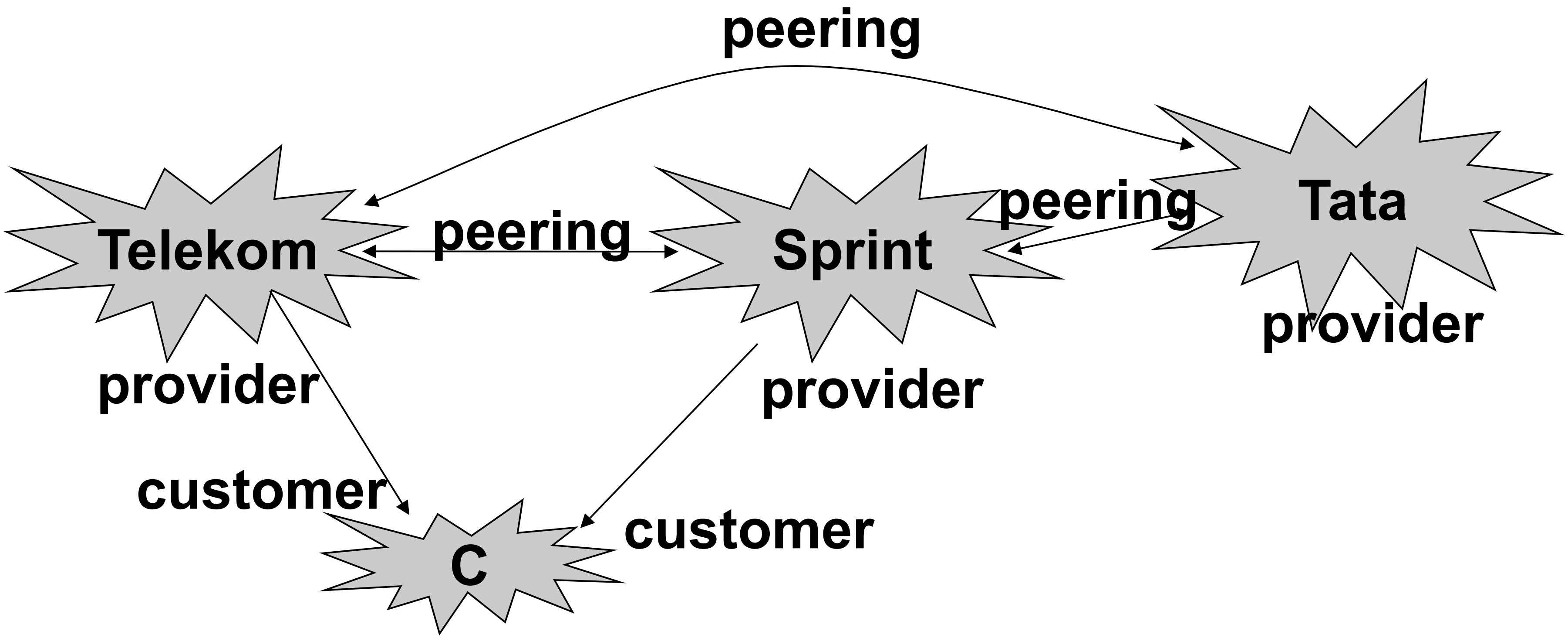
1. upstream: sequence of C→P links (possibly length = 0)
2. then possibly across *one* peering link
3. then downstream: sequence of P→C links (possibly length = 0)

But:
Sibling–sibling edges may occur at any position on a packet's path



Business and policy routing (10): "Tiers" / "DFZ"

- Big players have no providers, only customers and peers
  - "Tier-1" ISPs
  - or "Default-Free Zone" (DFZ)
    - have no default route to a "provider"
- Each Tier-1 peers with each other



Tier-1, Tier-2, Tier-3 etc.

- Tier-1/DFZ = only peerings, no providers
- Tier-2 = only peerings and one or more Tier-1 providers
- Tier-3 = at least one Tier-2 as a provider
- Tier-*n* = at least one Tier-(n-1) provider
  - defined recursively
  - $n \geq 4$: Rare in Western Europe, North America, East Asia

- "Tier-1.5" = almost a Tier-1 but pays money for *some* links
  - Example: Deutsche Telekom used to pay money to Sprint, but is now Tier-1
  - Marketing purposes: Tier-1 sounds better

Siblings

- Not everything is provider/customer or peering
- Sibling = mutual transit agreement
  - Provide connectivity to the rest of the Internet for each other
  - ≈ very extensive peering
- Examples
  - Two small ASes close to each other that do not want to afford additional Internet services
  - Merging two companies
    - Merging two ASes into one = difficult,
    - Keeping two ASes and exchanging everything for free = easier
  - Example: AT&T has five different AS numbers (7018, 7132, 2685, 2686, 2687)

BGP policy routing: Technical summary

1. Receive BGP update
2. Apply import policies
   - Filter routes
   - Tweak attributes (advanced topic…)
3. Best route selection based on attribute values
   - Policy: Local Pref settings and other attributes
   - Install forwarding tables entries for best routes
   - (Possibly transfer to Route Reflector)
4. Apply export policies
   - Filter routes
   - Tweak attributes
5. Transmit BGP updates

BGP policy routing: Business relationship summary

- Import Policy = Which routes to use
  - Select path that incurs most money
  - Special/political considerations (e.g., Iranian AS does not want traffic to cross Israeli AS; other kinds of censorship)
- Export Policy = Which routes to propagate to other ASes
  - Not all known routes are advertised:
    Export only…
    - If it incurs revenue
    - If it reduces cost
    - If it is inevitable
- Policy routing = Money, Money, Money…
  - Route import and export driven by business considerations
  - But *not* driven by technical considerations
    Example: Slower route via peer may be preferred over faster route via provider

Where to peer

(Here: Peering = having a BGP relationship)

A) Private peering

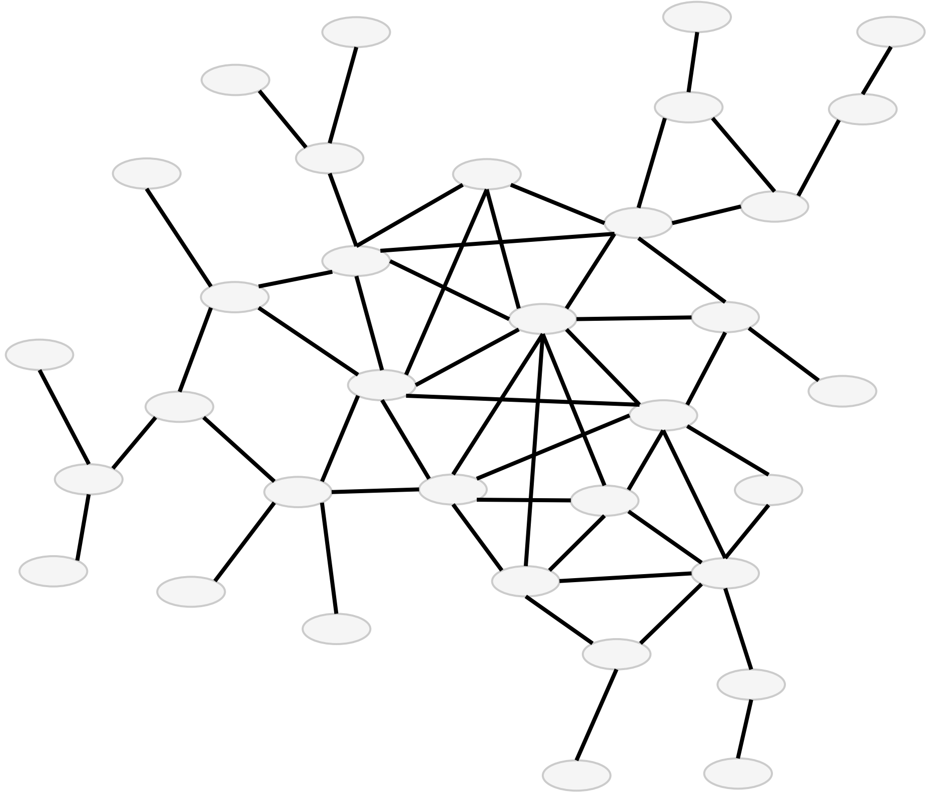
- The obvious solution: “Let’s have a cable from your server room to our server room”

B) At public peering locations (Internet Exchange Point, IX, IXP)

- “A room full of switches that many providers connect to”
- Configure VLAN connections in switch, instead of having to put in $O(n^2)$ separate wires
- Examples:
  - DE-CIX, Frankfurt (purportedly largest in world)
  - AMS-IX, Amsterdam
  - LINX, London
  - MSK-IX, Moscow

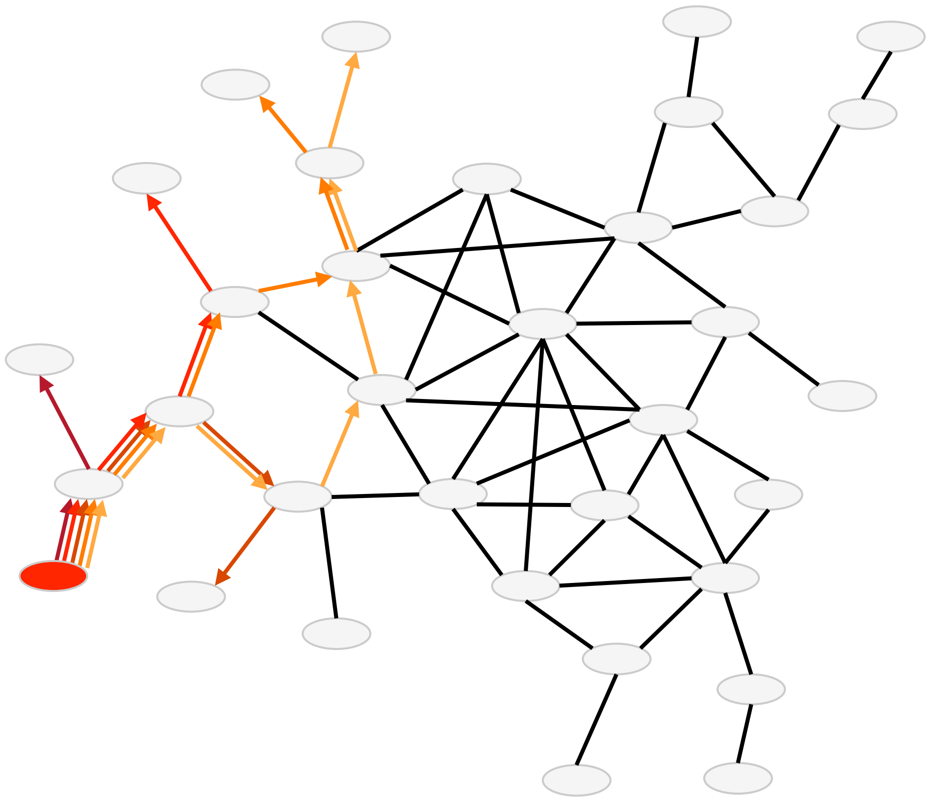
BGP Path Analysis

- Graph analysis
  - ASes as nodes
  - Links in AS path als edges
  - „Snapshot“ of Internet routes
  - Router-specific viewpoint



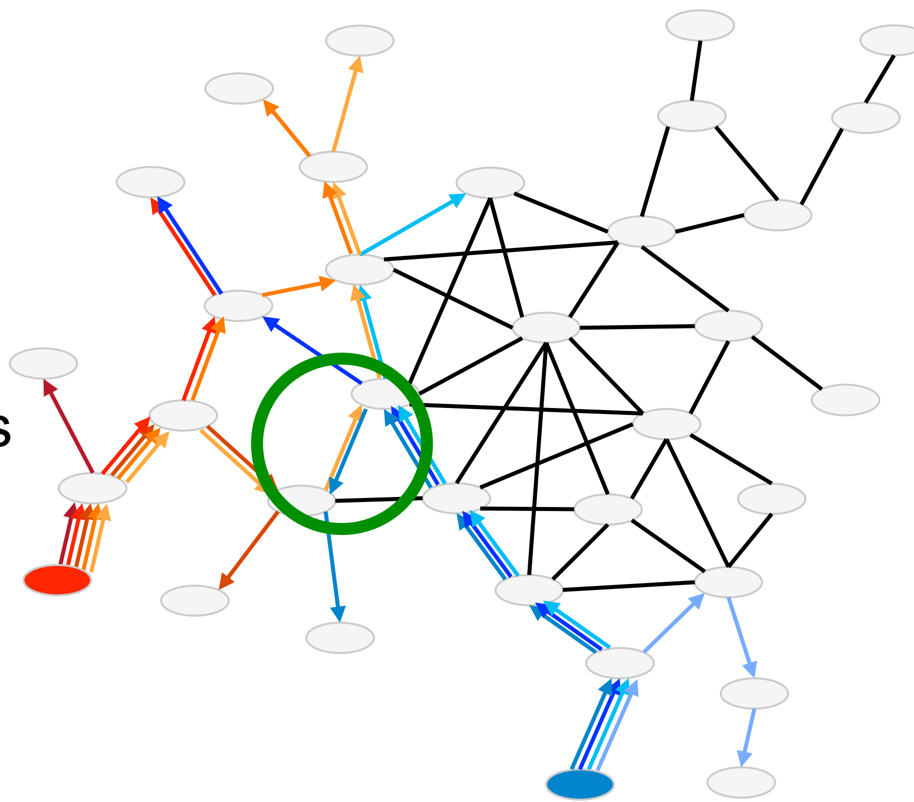
BGP Path Analysis

- Graph analysis
  - ASes as nodes
  - Links in AS path als edges
  - „Snapshot“ of Internet routes
  - Router-specific viewpoint



BGP Path Analysis

- Graph analysis
  - ASes as nodes
  - Links in AS path als edges
  - „Snapshot“ of Internet routes
  - Router-specific viewpoint
- Interesting nodes
  - large in- and out-degree
  - Internet fixpoints
- Route changes
  - observable in BGP updates
  - convergence prozess

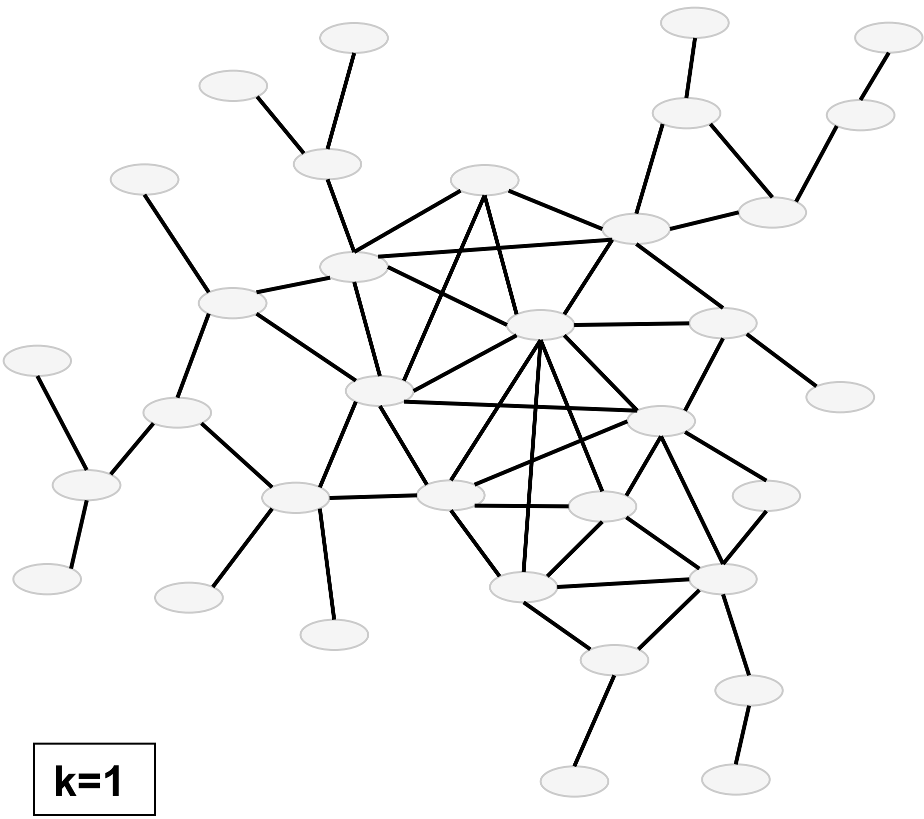


Internet Fixed Points

- Necessary properties of fixed point
  - Stable over long period of time
  - constant properties
  - Fixed point from different perspectives
  - Core as center of gravity: route length to fixed point is similar
- Candidates
  - Individual routers
  - Individual Autonomous System
  - Set of routers / Autonomous Systems
  - Structural components of Internet graph
- Core of the Internet
  - Set of Autonomous Systems
  - Stable (no significant fluctuation)
  - Fixed point from all perspectives
  - ⇨ k-core algorithm

Core of the Internet

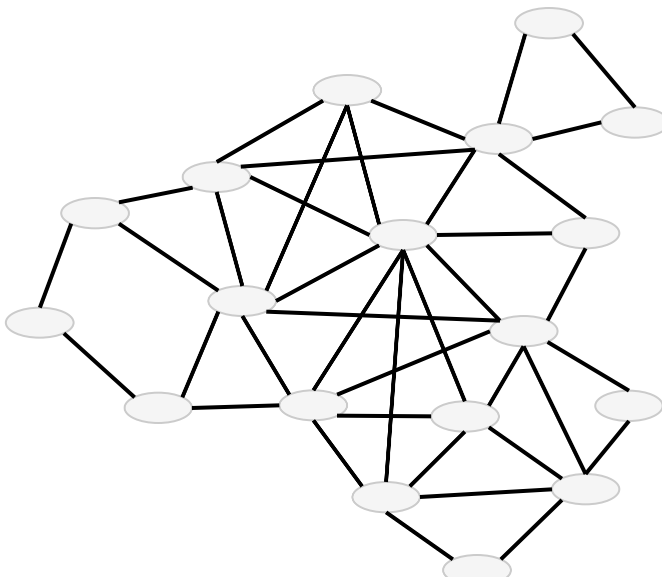
k-core algorithm



1. removal of nodes with degree=1

Core of the Internet

k-core algorithm



1. removal of nodes with degree=1
2. removal of nodes with degree<= 2

...

X. all nodes removed → (X-1)-core found

k=2

k-core algorithm

Internet AS core

- maximum k=23
- 49 AS (of 38.693 AS)

AS174 COGENT-174
AS209 ASN-QWEST
AS286 KPN
AS293 ESNET
AS701 UUNET
AS812 ROGERS-CABLE
AS852 UNKNOWN
AS1239 SPRINTLINK
AS1273 CW
AS1299 TELIANET
AS1668 AOL-ATDN
AS2497 Asia Pacific NIC
AS2516 KDDI
AS2828 XO-AS15
AS2914 NTT-COMM
AS3257 TINET-BACKBONE
AS3292 TDC
AS3303 SWISSCOM
AS3320 DTAG
AS3356 LEVEL3
AS3491 BTN-ASN
AS3549 GBLX
AS3561 SAVVIS
AS4134 APNIC
AS4323 TWTC
AS4436 AS-NLAYER
AS4637 REACH
AS5400 BT
AS5413 UNKNOWN
AS6453 UNKNOWN
AS6461 ABOVENET
AS6539 GT-BELL
AS6762 SEABONE-NET
AS6939 HURRICANE
AS7018 ATT-INTERNET4
AS7473 SINGTEL-AS-AP
AS8001 NET-ACCESS-CORP
AS8075 MICROSOFT-CORP
AS8928 INTEROUTE
AS9002 RETN-AS
AS10026 PACNET
AS10310 YAHOO-1
AS11164 TRANSITRAIL
AS13030 INIT7
AS15169 GOOGLE
AS15412 FLAG-AS
AS19151 WVFIBER-1
AS20940 AKAMAI-ASN1
AS22822 LLNW

BGP Update Process

- Neighboring node „announced“ route to destination prefix
  - Propagation of best route only
  - However: several routes to destination prefix known
  - Selection of best route as part of BGP Path Selection Process; influences include AS path length

- Evaluation
  - Statistical analysis (e.g. „number of route updates per prefix and time)
  - Quantitative analysi (e.g. number of topological changes of BGP graph)

- Convergence of BGP

BGP Update Process

- Example: process after route outage
  - Outage of link/system at destination D
  - Propagation of BGP messages
  - Convergence at observer O
- Process influenced by
  - BGP timeout (90s)
  - Number of different routes to destination
  - Withdrawal of all affected routes required for convergence