# How dynamic are IP Addresses?

Paper

## “How dynamic are IP Addresses?”

Y. Xie, F. Yu, K. Achan, E. Gillum, M. Goldszmidt, T. Wobber

Network Architectures: Internet Routing
WS 2007/08

Benjamin Vahl

# How dynamic are IP Addresses?



- Thematik des Papers?
  - Erkenntnisse über Charakteristiken dynamischer IP Adressen
  - Anwendung beim automatischen Identifizieren
  - Umsetzung: Algorithmus *UDmap*

# How dynamic are IP Addresses?



- IP-Blacklisting
  - SpamMail-Server, Botnet-Clients, Phishing-Sites ect. fast ausschliesslich von Rechnern hinter dynamischen IP-Adressen
  - Bisherige Gegenmaßnahme: Blacklisting-Dienste
  - Nachteile:
    - Listen nie aktuell
    - veralten schnell / Einträge werden redundant
    - unvollständig

# How dynamic are IP Addresses?



- Alternative
  - Automatisiertes Erkennen dynamischer IP-Blöcke durch Algorithmus
  - Blocken zugehöriger IPs

# How dynamic are IP Addresses?



- Theorie: Analyse dynamischer IPs
  - Eigenschaften
  - Rückschlüsse auf IP-(Block)-Nutzung
- Anwendung: *UDmap*
  - Voraussetzungen / Ergebnisse
  - Workflow

# How dynamic are IP Addresses?



## Eigenschaften dynamischer IP-Adressen

# How dynamic are IP Addresses?



- Trivial
  - Dynamische IP-Adresse ist nacheinander mehreren Benutzern zugeordnet
  - entsprechende Benutzer verwenden längerfristig unterschiedliche IPs

# How dynamic are IP Addresses?



- Auftreten in Blöcken
  - dynamische IP-Adressen fast immer Teil eines zusammenhängenden Bereichs (Pool)
  - Blöcke meistens Teilbereich eines IP-Prefixes
  - IP-Adressen innerhalb eines Blocks weisen ähnliche Charakteristiken auf
    ⇒ Rückschluss auf Zugehörigkeit

# How dynamic are IP Addresses?

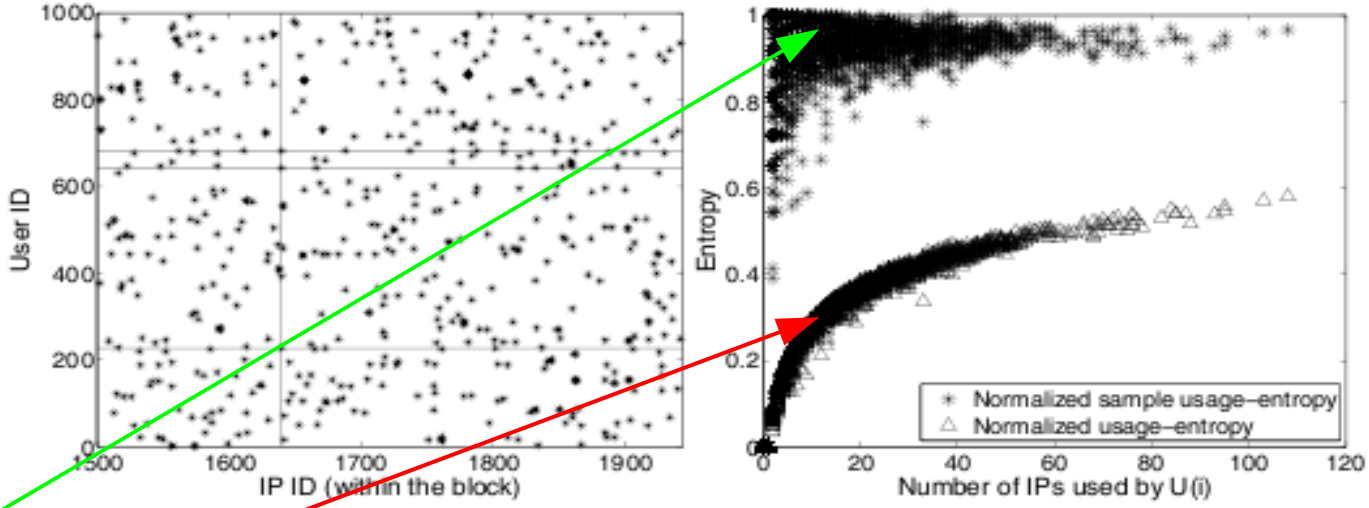


- IP-Benutzungsentropie
  - Benutzer bewegen sich mit ihren IPs innerhalb des Blocks
  - Wahrscheinlichkeit für Vergabe der IP innerhalb des Blocks gleichmäßig verteilt
  - Benutzungsentropie ≙ Wahrscheinlichkeitswert

# How dynamic are IP Addresses?



- IP-Benutzungsentropie
  - Beispiel




U(I): Gruppe von Benutzern mit $IP_i$

- U(i) erhält eine beliebige andere IP aus dem gesamten Block
- U(i) erhält beliebige IP des Blocks aus der Menge der tatsächlich insgesamt von U(i) verwendeten IPs

# How dynamic are IP Addresses?



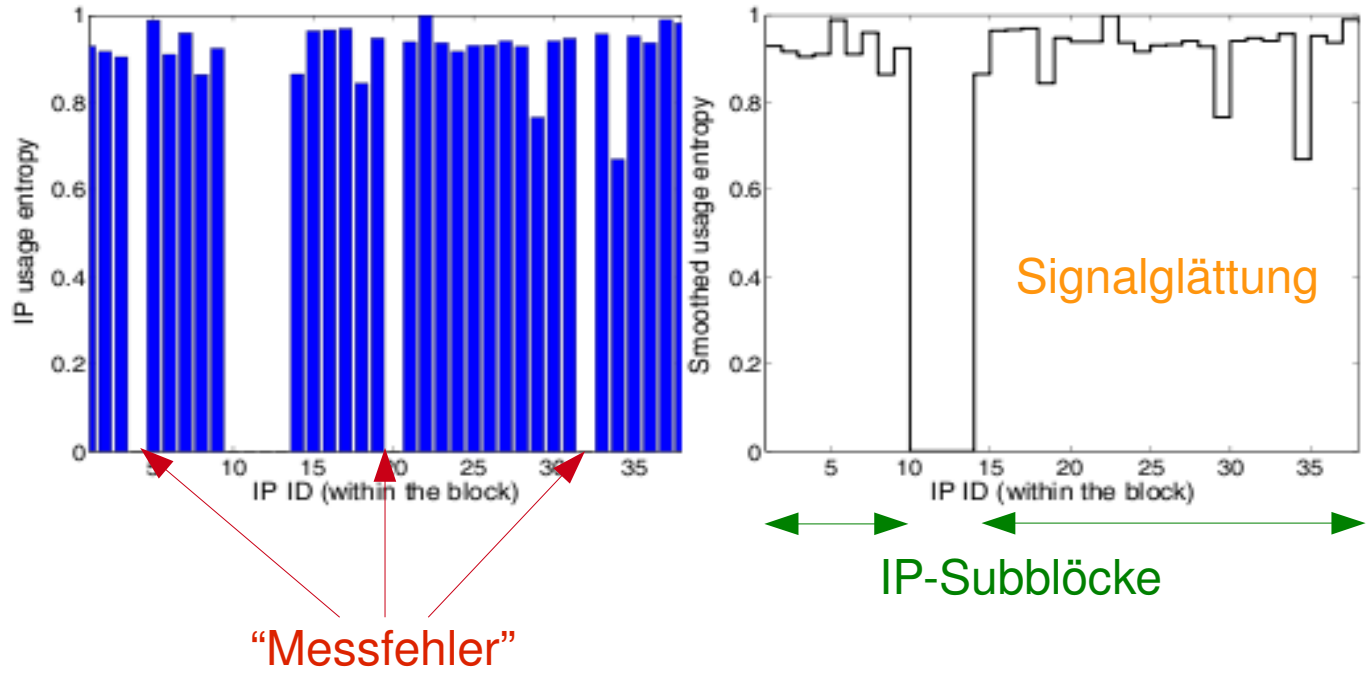
- Anwendung der Entropie
  - Entropiewert → 1 lässt auf dynamische IP schliessen
  - Zusammenfassung zu IP-Block erfordert Entropie der Mehrheit aller enthaltenen Adressen über Schwellenwert
  - Einsetzen von Signalglättung zur Korrektur von Ausreißern

# How dynamic are IP Addresses?



- Anwendung der Entropie
  - Beispiel

Signalglättung



# How dynamic are IP Addresses?



- IP-Volatilität
  - Maßeinheit für Rate des Benutzerwechsels einer dynamischen IP
  - Zeitraum gibt Aufschluß über Nutzungsart
  - mittlere Benutzungsdauer und Benutzer pro IP ausschlaggebend

# How dynamic are IP Addresses?

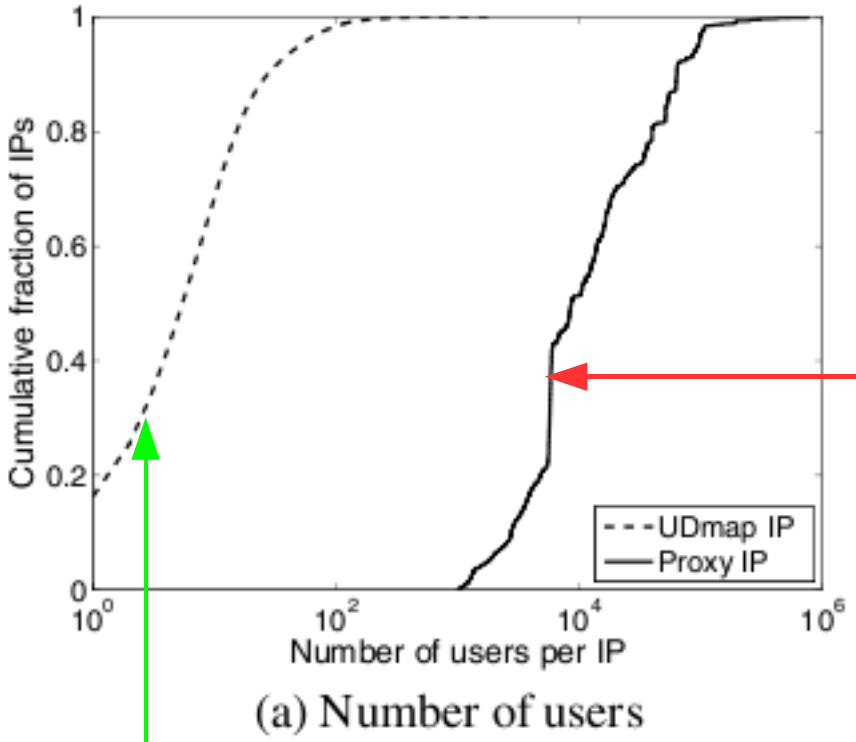


- IP-Volatilität
  - hilft, dynamische IPs von Proxies oder öffentlich genutzten Rechnern zu unterscheiden
    - sehr kurzfristig ⇒ Proxy
    - längerer Zeitraum ⇒ echte dynamische IP

# How dynamic are IP Addresses?



- Benutzer pro IP-Adresse




(a) Number of users

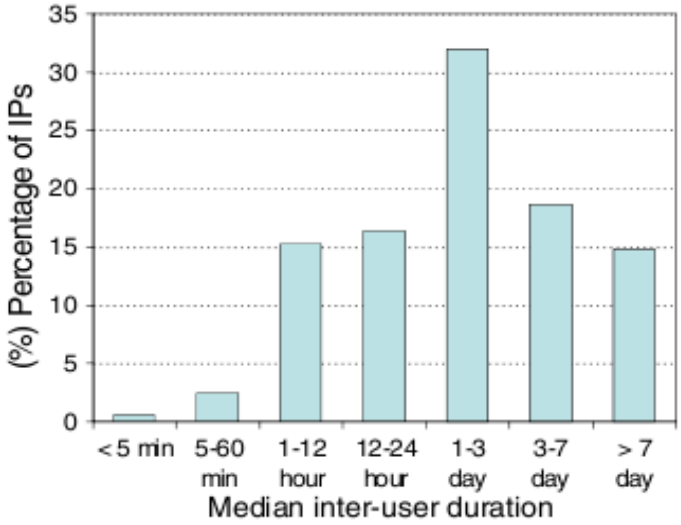
Proxy-Server, ect.
(false positives)

dynamische IP-Adressen

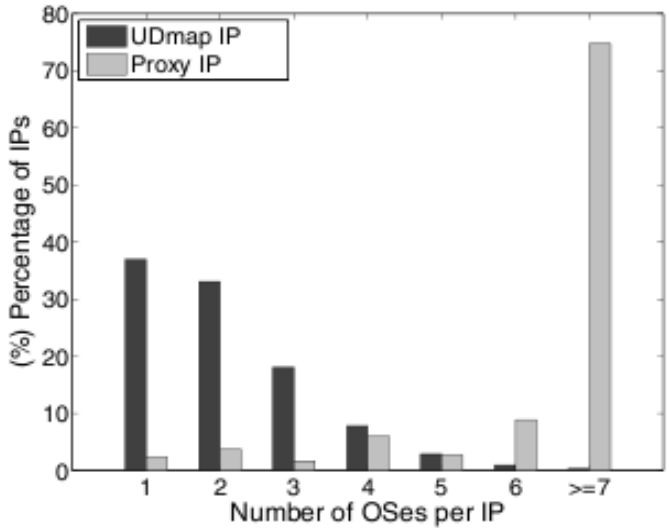
# How dynamic are IP Addresses?



- mittlere Benutzungsdauer




- Betriebssysteme pro IP



# How dynamic are IP Addresses?



- Streuungsfaktor
  - IPs innerhalb eines Blocks sollten erwartungsgemäß ähnliche Volatilitäts-eigenschaften besitzen
  - Streuungsfaktor quantifiziert Streuung der Volatilität, je größer desto stärker verteilt
  - ≙ normalisiertem Abstand zwischen 90% des Maximalwertes und dem Mittelwert

# How dynamic are IP Addresses?



- Erkenntnisse
  - Anzahl der Benutzer pro IP gleichmäßiger verteilt als die mittlere Benutzungsdauer
  - Kleinere Blöcke weisen meistens geringere Streuung auf, als größere

# How dynamic are IP Addresses?



- reverse DNS
  - Untersuchen des rDNS-Records bietet die einfache Informationsquelle über IP-Adressen zu treffen
  - Beispiel:
    IP: `157.57.215.19`
    Hostname: `adsl-dc-305f5.adsl.wanadoo.nl`
    ⇒ ADSL Anschluss in den Niederlanden

# How dynamic are IP Addresses?



- Probleme
  - Längst nicht alle Domains bieten für ihre IP-Adressen rDNS Records an
  - Einige rDNS-Namen sind nicht ausreichend informativ

# Umsetzung:
# Algorithmus UDmap

# How dynamic are IP Addresses?



- Voraussetzungen
  - Log-Dateien beliebigen Formats, die Host/Benutzer-IP-Zuordnung und ggf. Betriebssystem ermöglichen
  - Informationen über Netztopologie

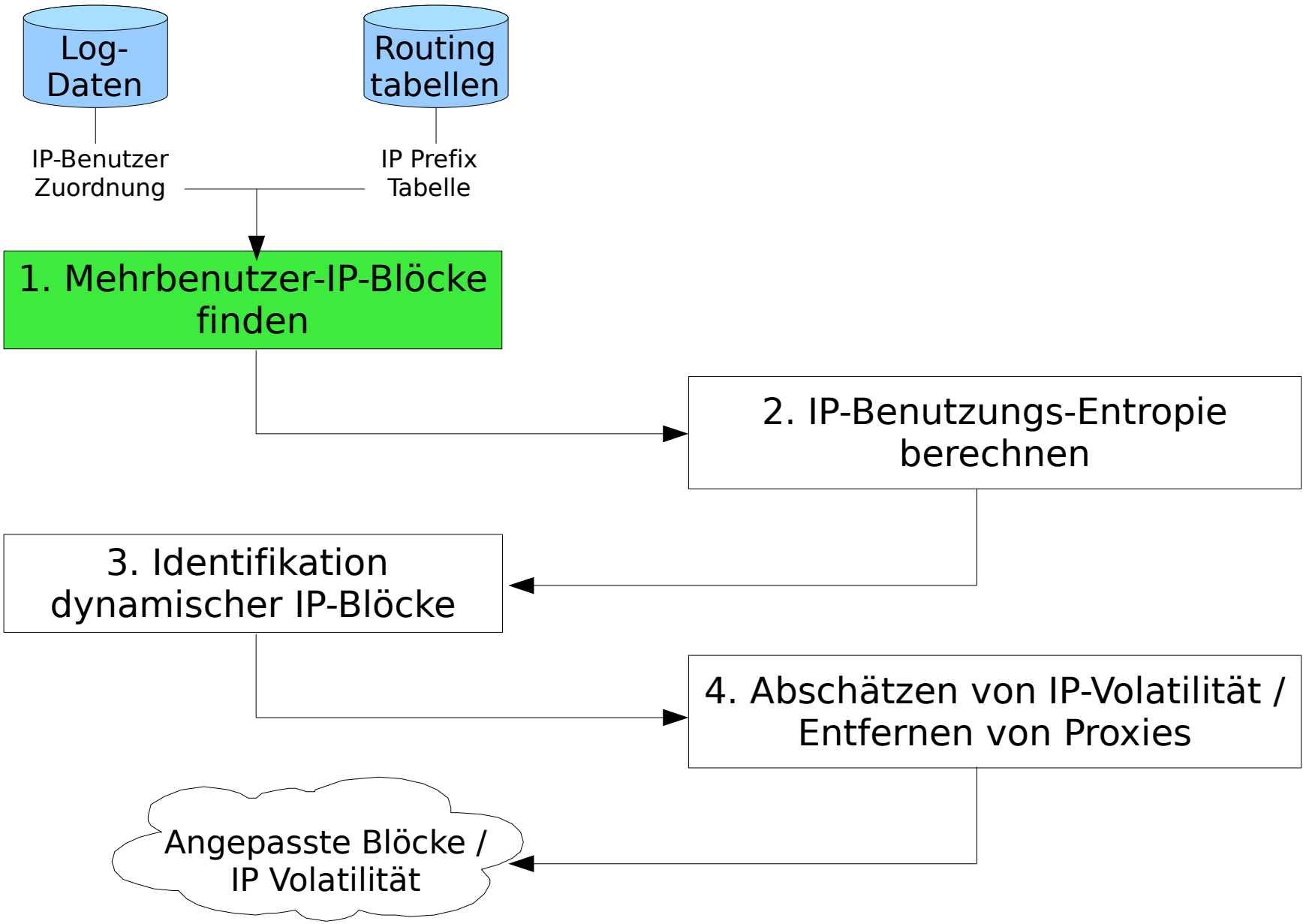
# How dynamic are IP Addresses?



- Ergebnisse nach Verarbeitung
  - möglichst präzise eingegrenzte dynamische IP-Blöcke
  - keine IPs von Proxy-Servern oder öffentlich zugänglichen Rechnern
  - IP-Volatilitätswerte

# How dynamic are IP Addresses?

Workflow ← UDmap



# How dynamic are IP Addresses?

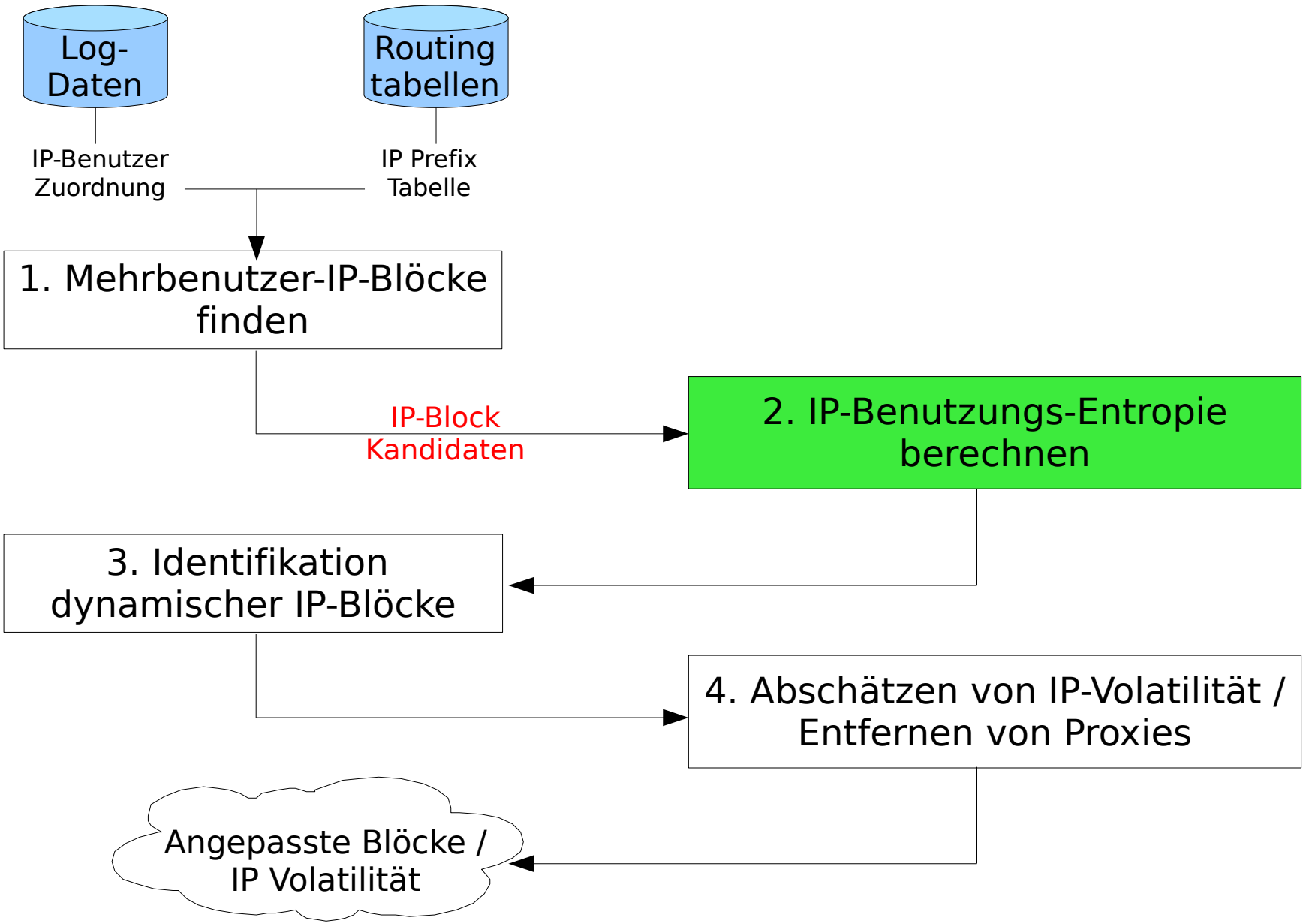


## 1. Mehrbenutzer-IP Blöcke finden

- mehrere aufeinanderfolgende IPs erscheinen dynamisch
- Kriterien:
  1. alle Adressen haben gleiches Routing-Prefix
  2. von $m$ IP-Adressen müssen mindestens $k$ Adressen im Log vorhanden und Lücken nicht größer als $g$ sein
  3. Anfangs- und Endadresse des Blocks müssen im Log auftreten

# How dynamic are IP Addresses?

Workflow ← UDmap



# How dynamic are IP Addresses?

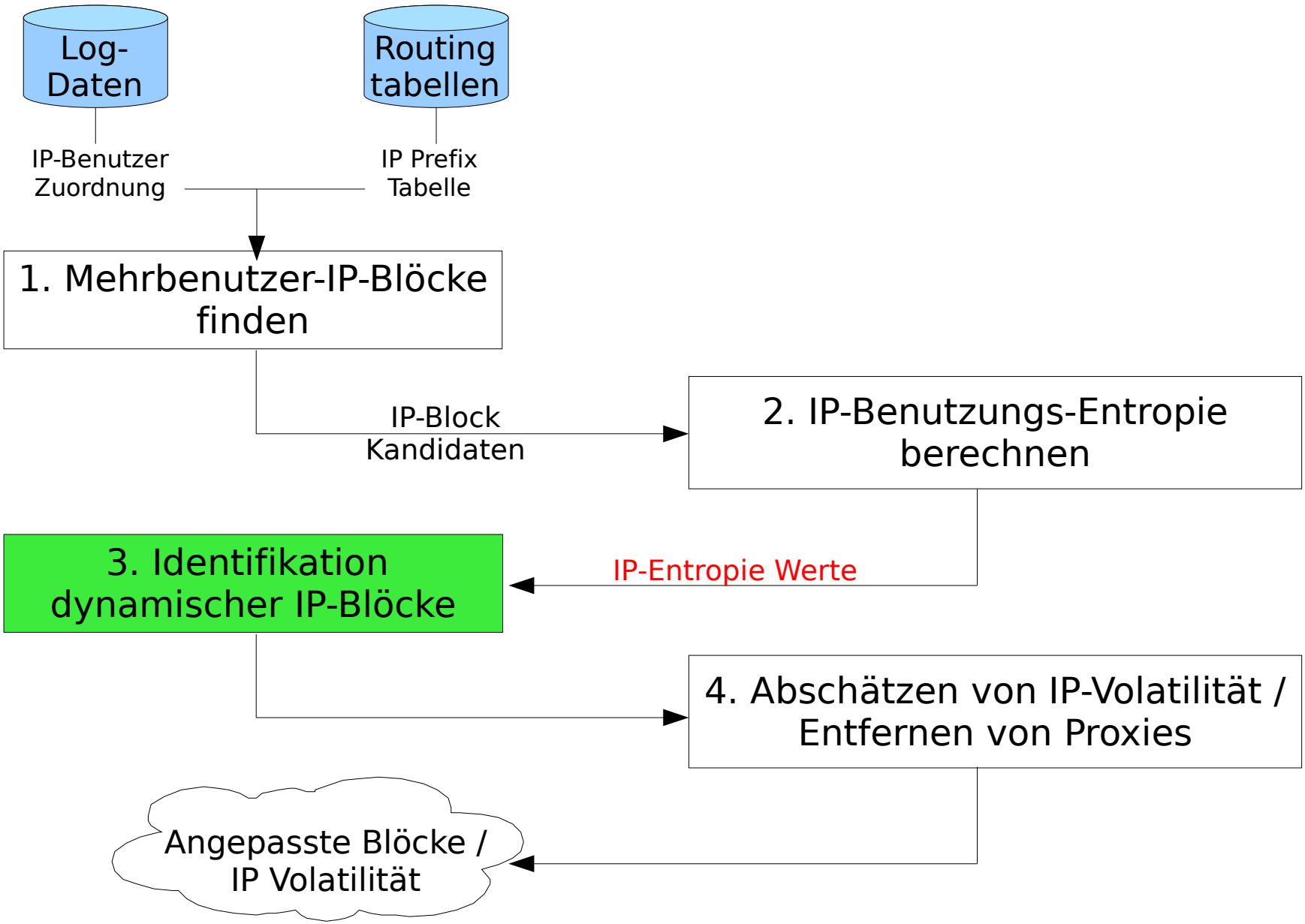


## 2. IP-Benutzungs-Entropie berechnen

- Ziel: Prüfen, ob einzelne IP-Adressen tatsächlich dynamisch und der Block zusammengehörig ist

# How dynamic are IP Addresses?

Workflow ← UDmap

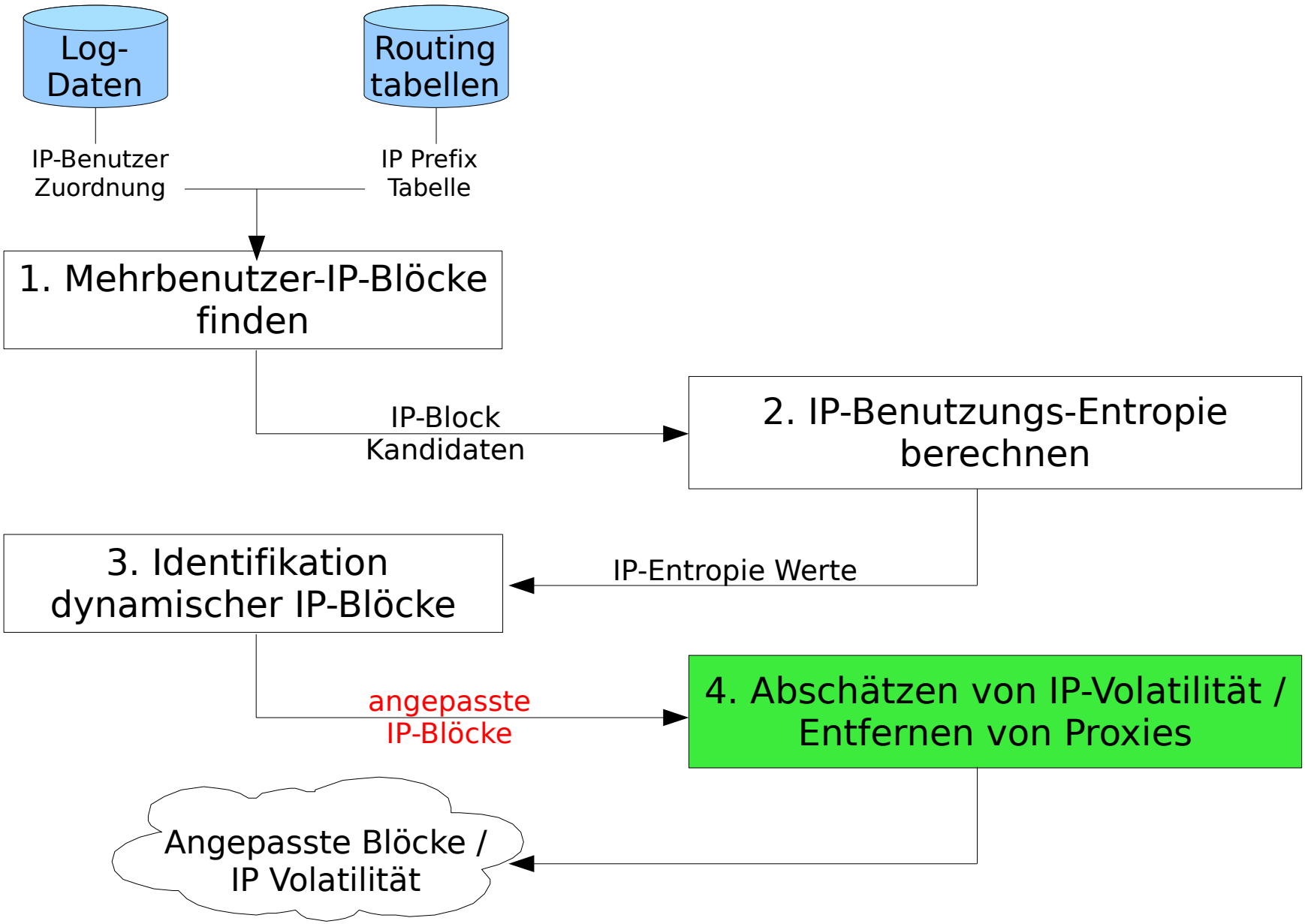


# 3. Identifikation dynamischer Blöcke

- Block-Kandidaten werden anhand Benutzungsentropie ggf. in Subblöcke unterteilt
- Anwenden der Signalglättungsfunktion

# How dynamic are IP Addresses?

Workflow ← UDmap



# How dynamic are IP Addresses?



## 4. IP-Volatilität, Entfernen von Proxys

- Volatilitätswerte von IP-Adressen innerhalb der jetzt angepassten Blöcke berechnen
- darüber identifizierte Proxies ect. ausschliessen

# How dynamic are IP Addresses?



- Validierung der Ergebnisse
  - Abgleich mit Dynablock Blacklisting Tabellen
  - verbleibende Adressen werden über rDNS verifiziert

# How dynamic are IP Addresses?



- Revision
  - ohnehin schon gefragtes netzwerk-basiertes Filtern wird noch effizienter
  - könnte langfristig Mailprovider unabhängiger von Blacklisting-Diensten machen
  - Kostenersparnis durch weniger Traffic
  - Erkenntnisse über Charakteristik dynamischer IPs nicht nur für Blacklisting sondern auch Forschung interessant

# How dynamic are IP Addresses?



- Quellen
  - How dynamic are IP Addresses?
    Y. Xie, F. Yu, K. Achan, E. Gillum, M. Goldszmidt, T. Wobber
    ACM SIGCOMM 2007
  - In Conference on Email and Anti-Spam
    A. Ramachandran, D. Dagon, and N. Feamster
    ACM 2006

# How dynamic are IP Addresses?

O RLY?

Fragen?

(YA RLY!)

# How dynamic are IP Addresses?



## 2. IP-Benutzungs-Entropie ...

- Jedem IP-Block werden für die gesamte Menge an Benutzern U aus dem Log die Benutzungen markiert:

  $A \in \{0,1\}^{|U| \times m}$ zugeordnet (Adjazenzmatrix)

- Benutzungs-Entropiewert $H(j)$ repräsentiert für alle Benutzer $U(j)$ die Wahrscheinlichkeit, eine andere Adresse aus diesem Block zu erhalten

# How dynamic are IP Addresses?



## 2. IP-Benutzungs-Entropie ...

- $H(j)$ wird für jeden Benutzer aus $U(j)$ berechnet und wegen der unterschiedlichen Blockgröße, auf einen Wert zwischen 0 und 1 normalisiert
- Dabei wird unterschieden zwischen
  - $H_B(j)$, normalized usage entropy
  - $H_U(j)$, normalized sample usage entropy

# How dynamic are IP Addresses?



## 2. IP-Benutzungs-Entropie ...

- $H_B(j)$ gibt an, ob die Verteilung der

  Wahrscheinlichkeit der Benutzer $U(j)$, eine andere IP des Blocks zu erhalten, gleichmäßig verteilt ist ($H_B(j) \rightarrow 1$ für hohe Wahrscheinlichkeit)

  $H_U(j)$ ähnlich wie $H_B(j)$, beschränkt sich aber

  **ausschliesslich** auf die von $U(j)$ im Block verwendeten Adressen

# How dynamic are IP Addresses?



## 2. IP-Benutzungs-Entropie ...

- Idealfall: Adressen aus dem betrachteten Block werden zufällig vergeben, $H_B(j)$ ist für die meisten Adressen näherungsweise 1
- In der Realität meistens nicht der Fall, da der im Log betrachtete Zeitraum zu kurz ist, um realistische Verteilung zu erhalten
- Bei begrenzter Datenmenge schätzt $H_U(j)$ die Gleichverteilung der Adressauswahl besser ab