Algorithmen für Ad-hoc- und Sensornetze

VL 08 – Data Gathering mit Network Coding

Dr. rer. nat. Bastian Katz

Lehrstuhl für Algorithmik I
Institut für theoretische Informatik
Universität Karlsruhe (TH)
Karlsruher Institut für Technologie

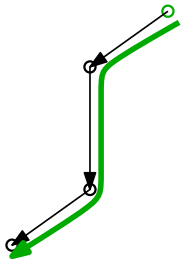
17. Juni 2009
(Version 3 vom 22. Juni 2009)

Heute

- Network Coding, eine Einführung
  - Ursprung in drahtgebundenen Netzen
  - Anwendung in drahlosen Netzen
- Data Gathering + Network Coding
  - Vergleich zur letzten Vorlesung
  - Zwei Ansätze zur Minimierung der Energie

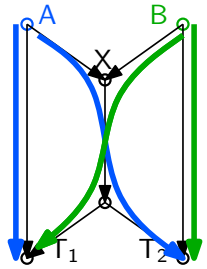
Informationsaustausch bisher..

- Daten werden von (einem/vielen) Knoten an (einen/viele) andere geschickt.
- Pakete werden
  - abgeschickt
  - weitergeleitet
  - empfangen



Informationsaustausch bisher..

- Daten werden von (einem/vielen) Knoten an (einen/viele) andere geschickt.
- Pakete werden
  - abgeschickt
  - weitergeleitet
  - empfangen



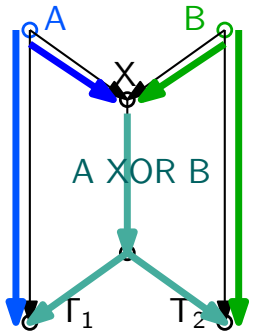
Wenn A und B jeweils N Pakete für T_1 und T_2 haben, wie viele Zeitschritte dauert es, diese zuzustellen, wenn über jede gerichtete Kante in jeder Zeiteinheit ein Paket übertragen werden kann?

- Klassische Antwort: etwas über 2N Zeitschritte!
  - Mittlere Kante muss jedes Paket einmal übertragen!

..und mit Network Coding

Das geht besser!

- A und B schicken nacheinander alle Pakete an X_1 und an T_1 bzw. T_2
- X_1 verknüpft ankommende Pakete von A und B per *XOR* und schickt sie an X_2
- X_2 schickt die XOR-Pakete an T_1 und T_2
- fertig in $N + 3$ Zeitschritten!

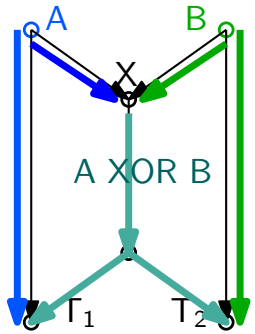


Klassisches Beispiel: Vermischen von Datenflüssen erhöht Durchsatz in drahtgebundenen Netzen!

Network Coding

Definition

Network Coding bezeichnet alle Techniken, bei denen die Pakete verschiedener Informationsströme miteinander vermischt werden.

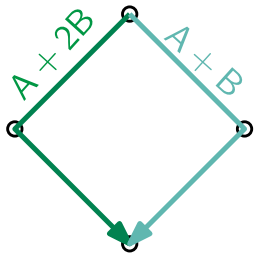


Network Coding

Definition

Network Coding bezeichnet alle Techniken, bei denen die Pakete verschiedener Informationsströme miteinander vermischt werden.

- Konsequenz: einzelne Pakete werden in der Regel wertlos!
  - das kann auch ein Feature sein!
  - Angreifer müssen mehrere Datenströme abgreifen

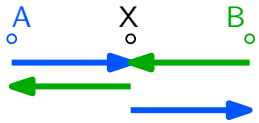


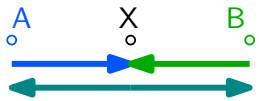
Network Coding

Definition

Network Coding bezeichnet alle Techniken, bei denen die Pakete verschiedener Informationsströme miteinander vermischt werden.

- Konsequenz: einzelne Pakete werden in der Regel wertlos!
  - das kann auch ein Feature sein!
  - Angreifer müssen mehrere Datenströme abgreifen
- Hilft das auch in drahtlosen Netzen?
  - Ja, das kann Übertragungen sparen!
  - Beispiel: Austausch zwischen zwei Knoten



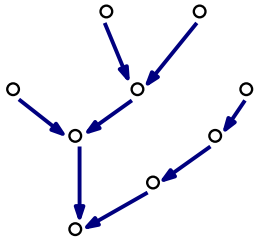


Data Gathering bisher...

I) (Fast) vollständige Aggregation

Jeder Knoten empfängt Pakete seiner Kinder und sendet *ein* Paket Richtung Senke

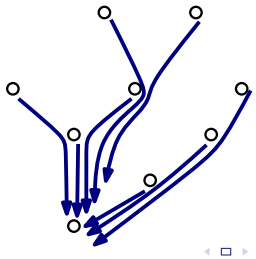
- ≫ Distributive, Algebraische Funktionen
- ≫ sogar Median funtioniert so
  - ≫ (in vielen Runden, gepipelined...)



II) Data Gathering ohne Aggregation

Jeder Knoten schickt sein(e) Paket(e) ab und leitet Pakete aus gesamtem Teilbaum weiter.

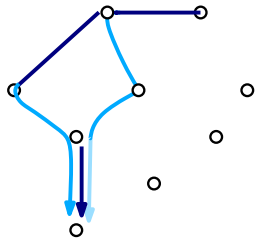
- ≫ typisch für wirklich komplexe Anfragen oder dann, wenn alle Daten zur Analyse benötigt werden.



...und jetzt viel allgemeiner

III) Network Coding

Jeder Knoten v_i hat Paket p_i mit $|p_i|$ Bits. Er darf es verschicken, Knoten können es umkodieren, am Ende muss die Senke das Paket rekonstruieren können.



- ≫ vollständige Aggregation fällt eig. nicht darunter
  - ≫ Senke erhält nicht alle Daten
- ≫ DG ohne Aggregation ist Sonderfall
  - ≫ Pakete werden nie umkodiert und auf Baum zur Senke geroutet
- ≫ wir betrachten das Problem noch allgemeiner, nämlich nicht nur beschränkt auf Bäume!

Optimale Topologien

Erinnerung

Wir sind bei der Aggregation und dem Einsammeln von Daten bisher davon ausgegangen, dass ein Baum gegeben ist, auf dem die Pakete zur Senke geroutet werden.

Was, wenn wir nur einen Graphen vorgeben und ganz individuell für Pakete angeben können, wie sie weitergereicht werden und wo sie umkodiert werden sollen?

- ≫ Kosten c_e für jede Kante e pro übertragenem Bit gegeben
- ≫ Ziel: Minimale Gesamtkosten

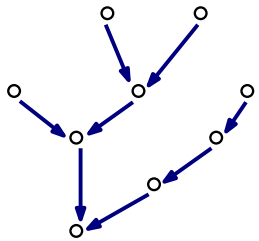
$$\sum_e c_e \sum_{p \text{ nutzt } e} |p|$$

Optimale Topologien (Warm-up I)

Beobachtung I

Bei uniformen Paketgrößen ist ein zur Senke gerichteter MST im Fall vollständiger Aggregation optimal.

- ≫ jedes Paket kostet nur auf erster Kante
- ≫ erste Kanten bilden gerichteten Spannbaum
- ≫ Routingkosten entsprechen genau den Kosten des Baumes



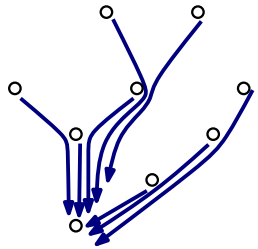
Optimale Topologien (Warm-up II)

Beobachtung II

Unabhängig von den Paketgrößen ist ein zur Senke gerichteter Kürzeste-Wege-Baum[a] beim DG ohne Aggregation optimal.

- ≫ jedes Paket „zahlt“ für Pfadlänge
- ≫ KW-Baum offensichtlich optimal
- ≫ zumindest zentral leicht zu berechnen

[a] bzgl. der Kosten



Schon wieder ist ein Baum optimal, das setzen wir aber nicht voraus! Pakete dürfen sich beliebig bewegen!

Data Gathering + Network Coding

Idee

Man kann auch alle Daten einsammeln und die Korrelationen zwischen den Daten ausnutzen!

- ≫ Zwei ähnliche Datenpakete nehmen weniger Platz ein, wenn man das erste schickt und statt des zweiten nur die Differenz/Quotienten usw.

Wir gehen davon aus, dass jeder Knoten v_i in jeder Runde ein Datenpaket mit $|p_i|$ Bits „produziert“. Wie kann man geschickt erreichen, dass eine Senke alle Pakete rekonstruieren kann, wenn wir Network-Coding-Techniken nutzen?

Time Coding

Nicht sehr einfallsreich: Time Coding

Knoten nutzen vorangegangene *eigene* Messungen, um Pakete zu kodieren

- ≫ Messwerte verändern sich nicht drastisch
- ≫ verschicke ab und zu komplette Messwerte und sonst nur Unterschiede
- ≫ (wie bei Videokompression)

Das ist *kein* echtes Network Coding! Network Coding nutzt Redundanzen über Knotengrenzen hinweg!

Multi-Input- vs. Single-Input-Coding

Grundprinzip

Ein Paket p_i eines Knotens v_i kann zu einem (kleineren) Paket $p_i^{j_1,j_2,\dots}$ kodiert werden, wenn Kodierer und Senke Kenntnis über Pakete $p_{j_1}, p_{j_2}, \dots$ haben.

Multi-Input-Coding

Daten eines Knoten können in Abhängigkeit der Daten *vieler* anderer Knoten kodiert werden

- \+ nutzt Redundanzen vieler Knoten aus
- \- setzt geordneten Informationsfluss voraus

Single-Input-Coding

Daten eines Knotens werden nur abhängig von Information *eines* anderen Knotens kodiert.

- \- nutzt weniger Redundanz
- \+ leichter zu koordinieren
- \+ jedes Paket wird nur einmal (um)kodiert

Single-Input-Coding: Annahmen

Beim Single-Input-Coding darf jedes Paket p_i nur einmal durch Ausnutzung eines Paketes p_j umkodiert werden. Wir

- ≫ bezeichnen dann p'_i als p_i^j
- ≫ gehen davon aus, dass wir für alle i, j wissen, wie stark sich p_i unter Kenntnis von p_j komprimieren lässt, also $|p_i^j|/|p_i|$

Achtung! Wer immer p_i zu p_i^j umkodiert, muss p_i weiterleiten und p_j zumindest rekonstruieren können.

Self-Coding vs. Foreign-Coding

Foreign-Coding

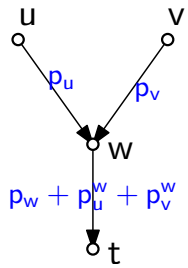
Jeder Knoten darf bisher unkomprimierte Daten, die er weiterleitet kodieren und dabei eigene Daten verwenden.

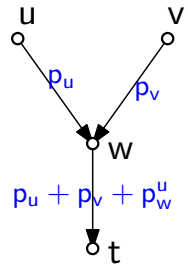
- ≫ Kodierung von p_i zu p_i^j in Knoten v_j!

Self-Coding

Jeder Knoten darf nur seine eigenen Daten kodieren und dabei Daten von Knoten verwenden, die er weiterleitet.

- ≫ Kodierung von p_i zu p_i^j in Knoten v_i
- ≫ das impliziert ..



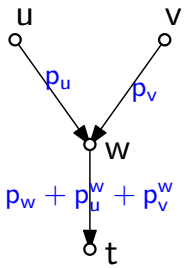


Problem: Single-Input Foreign-Coding

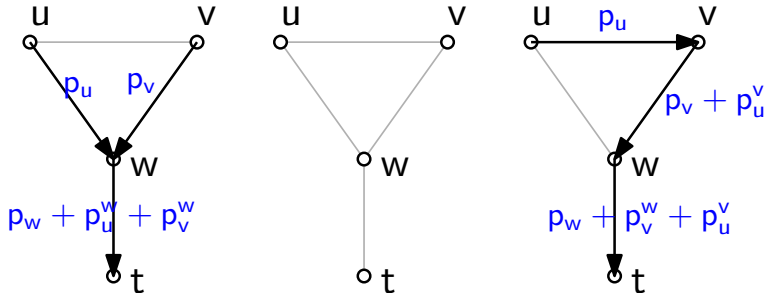
Gegeben Graph $G = V, E$, Kantenkosten c_e, Paketgrößen $|p_i|$ für alle $v_i \in V$, abhängige Paketgrößen $|p_i^j|$ für alle $v_i, v_j \in V$

Gesucht Routing und Kodierungsknoten darauf, so dass die Gesamtkosten minimiert werden

- ≫ *Wohin* wird ein Paket geschickt?
- ≫ *Wo* wird es umkodiert?
- ≫ *Wohin* wird das umkodierte Paket geschickt?



Ein kleines Beispiel



Unterschiedliches Routing kann zu unterschiedlichen Kosten führen!

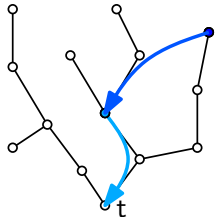
- ≫ Im Beispiel für einheitliche Kantenkosten
- ≫ links: $|p_u| + |p_v| + |p_w| + |p_u^w| + |p_v^w|$ (besser für $|p_u^w| < 2|p_u^v|$)
- ≫ rechts: $|p_u| + |p_v| + |p_w| + 2|p_u^v| + |p_v^w|$ (besser sonst)

Beobachtungen

Beobachtung I

Sobald ein Paket umkodiert wurde, sollte es auf dem billigsten Weg zur Senke geroutet werden

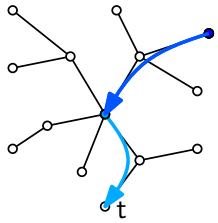
≫ braucht Kürzeste-Wege-Baum zu t



Beobachtung II

Um zu dem Knoten zu kommen, an dem es umkodiert wird, sollte ein Knoten idealerweise auf billigstem Weg geroutet werden.

≫ braucht alle Kürzeste-Wege-Bäume



≫ Wie wählen wir Kodierungsknoten?

Optimale Route

Zusammenfassung

Wird ein Paket p_i in einem Knoten v_j zu p_i^j kodiert, sind die Kosten für p_i minimal, wenn

≫ p_i es auf kürzestem Weg von v_i zu v_j geroutet wird

≫ p_i^j auf kürzestem Weg von v_j zur Senke t geroutet wird

≫ Kosten für Paket p_i sind dann

$$C(i,j) := |p_i| KW(v_i, v_j) + |p_i^j| KW(v_j, t)$$

(unkodierte Pakete haben Kosten $C(i,t)$)

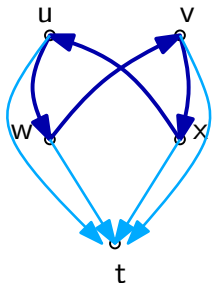
Trugschluss

Wenn man für jedes p_i das v_j zum Umkodieren so wählt, dass $C(i,j)$ minimiert wird, ist das Problem optimal gelöst.

Uhh-ohh

Definition

Der *Abhängigkeitsgraph* eines Single-Input-Codierungsschemas bezeichnet den gerichteten Graphen auf der Knotenmenge, der eine Kante (v_i, v_j) enthält, wenn p_i mit Hilfe von p_j kodiert wird.



Uhh-ohh

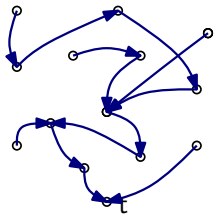
Definition

Der *Abhängigkeitsgraph* eines Single-Input-Codierungsschemas bezeichnet den gerichteten Graphen auf der Knotenmenge, der eine Kante (v_i, v_j) enthält, wenn p_i mit Hilfe von p_j kodiert wird.

Lemma

Die Senke kann die Pakete p_i nur dann rekonstruieren, wenn der Abhängigkeitsgraph zyklenfrei ist.

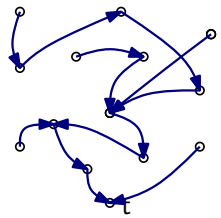
≫ wird p_i mit Hilfe von p_j kodiert, muss die Senke p_j zuerst rekonstruiert haben!



Optimale Lösung

Satz

Ist $T = (V, E)$ ein zur Senke gerichteter Baum, der
$$C_T := \sum_{(v_i, v_j) \in E} C(i, j)$$
minimiert, dann ist es optimal, wenn für jedes $(v_i, v_j) \in E$ das Paket p_i von v_j kodiert wird.



- ≫ $C(i, j)$ sind die Kosten, die Paket p_i erzeugt, entstehen, wenn es in v_j kodiert wird
  - ≫ wieder: $C(i, t)$ für unkodierte Pakete!
- ≫ die Kosten dieser Lösung sind genau C_T
- ≫ keine Lösung kann besser sein:
  - ≫ jede Lösung muss azyklisch sein
  - ≫ jedes Paket wird entweder irgendwo kodiert oder kommt unkodiert in t an!

MEGA: Minimum Energy Gathering Algo

1. Berechne Kürzeste-Wege-Bäume zu allen Knoten
   - ≫ (das ist einfach, aber viel Arbeit)
2. Berechne alle $C(i, j)$
3. Berechne DMST (gerichteten MST) $T = (V, E)$ mit Wurzel t und Gewichten $C(i, j)$
   - ≫ das geht (nicht in dieser Vorlesung)
4. route Paket p_i über v_j für $(v_i, v_j) \in E$
   - ≫ route p_i auf kürzestem Weg zu v_j
   - ≫ kodiere dort p_i zu p_i^j
   - ≫ route p_i^j auf kürzestem Weg zu t

Satz (Beweis ist schon vorbei)

MEGA löst das Single-Input-Foreign-Coding-Problem optimal!

Diskussion: MEGA verteilt?

- woher kommen $|p_i^j|/|p_i|$ für entfernte Knoten?
- benötigt Kürzeste-Wege-Bäume *aller Knoten*
- berechnet DMST auf vollständigem Graphen

Beobachtung

Wenn wir die Auswahl der Kodierungsknoten auf die Nachbarschaft einschränken, haben wir diese Probleme nicht!

+ benachbarte Knoten können können Korrellation ermitteln
+ *ein* Kürzeste-Wege-Baum von Senke reicht aus
+ Nur tatsächlich vorhandene Kanten sind relevant für DMST-Berechnung (das geht verteilt, ohne Beweis)

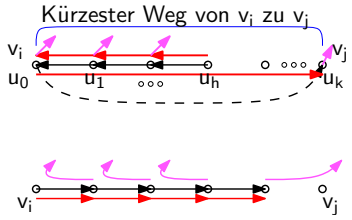
Und wir verlieren nichts! (fast)

Satz

Sind die Pakete aller Knoten gleich groß und nimmt $|p_i^j|/|p_i| = |p_j^i|/|p_j|$ mit der Entfernung zwischen v_i und v_j zu, dann findet MEGA in UDGs auch eine optimale Lösung, wenn nur benachbarte Knoten als Kodierungsknoten ausgewählt werden dürfen.

- ≫ Es reicht zu zeigen: Es gibt immer eine optimale Lösung, in der jedes Paket von einem benachbarten Knoten kodiert wird

- betrachte Knoten v_i mit Kodierungsknoten v_j außerhalb der Nachbarschaft und kürzesten Weg dorthin
  - es gibt ein erstes u_h, das keinen Pfad im DMST zu v_i hat

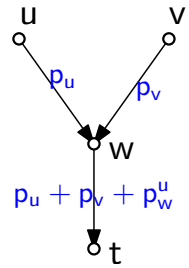


- Kosten der Pakete von u_0 bis u_h bis zur Kodierung $\left(2\sum_{i=1}^{h} c_{(u_{i-1},u_{i-1})} + \sum_{i=h+1}^{k} c_{(u_{i-1},u_{i-1})}\right)|p|$
- Ändere Kanten wie in Abbildung
- Kosten der Pakete von u_0 bis u_h bis zur Kodierung jetzt noch $\left(\sum_{i=1}^{h+1} c_{(u_{i-1},u_{i-1})}\right)|p|$
- Dazu im schlimmsten Fall noch Zusatzkosten um wieder genausoviele kodierte Pakete in allen Knoten zu haben wie vorher (falls das besser war) $\left(\sum_{i=1}^{h} c_{(u_{i-1},u_{i-1})}\right)|p| + \left(\sum_{i=h+2}^{k} c_{(u_{i-1},u_{i-1})}\right)|p|$
- Nachrechnen, das ist dasselbe!

Problem: Single-Input Self-Coding

Erinnerung: Self-Coding

- Paket p_i wird in v_i kodiert mit Hilfe eines Paketes p_j, das v_i rekonstruieren kann.



Gegeben Graph $G = V, E$, Kantenkosten c_e, Paketgrößen $|p_i|$ für alle $v_i \in V$, abhängige Paketgrößen $|p_i^j|$ für alle $v_i, v_j \in V$

Gesucht Routing und Kodierungszuordnung, die die Gesamtkosten minimieren

- *Wohin* werden Pakete verschickt?
- *Welches* Paket wird zur Kodierung verwendet?

Lösungen müssen nicht so einfach aussehen!

Satz (ohne Beweis)

Single-Input Self-Coding ist schon NP schwer für einheitliche Paketgrößen $|p_i| \equiv s$ und von der Quellen unabhängige Kompression $p_i^j \equiv s_e$.

- nur für diesen Fall gibt es einen Faktor-2-Approximationsalgorithmus
  - LEGA: Low-Energy Gathering Algorithm
  - weicht *wesentlich* vom eigentlichen Routingschema ab!
  - (eigentlich: Faktor $2(1+\sqrt{2})$ u. Kommunikation nur auf *einem* Baum)

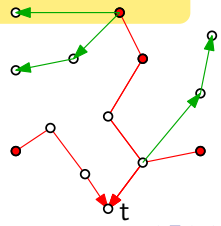
Beobachtungen

DG mit Self-Coding kostet mindestens so viel wie es kostet, alle kodierten Pakete auf dem billigsten Weg zur Senke zu bringen:

$$C_{OPT} \geq s_e \sum_{u \in V} KW(u,t)$$

Bei DG mit Self-Coding kostet mindestens so viel wie es kostet, alle kodierten Pakete auf dem billigsten Weg zur Senke zu bringen:

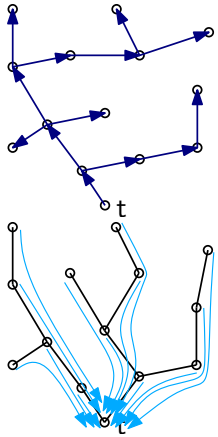
$$C_{OPT} \geq s \cdot c(MST)$$



- einige Knoten senden Rohdaten an die Senke
  - das ergibt Menge von Pfaden zwischen diesen Knoten und der Senke
- auch ein Knoten, der nicht auf einem dieser Pfade liegt, empfängt trotzdem mind. einmal Rohdaten!
  - das „verbindet“ sie mit den Pfaden

LEGA, vereinfacht

1. Berechne MST *und* Kürzeste-Wege-Baum
2. jedes v_i schickt sein p_i an seine Kinder im MST
3. jedes v_i kodiert sein p_i mit dem Paket seines Vorgängers im MST und schickt es im Kürzeste-Wege-Baum Richtung Senke



- Abhängigkeiten sind wieder azyklisch!
  $C_{\text{LEGA}} \leq s \cdot MST + s_e \cdot \sum_{u \in V} KW(u,t) \leq 2C_{\text{OPT}}$
- aber: kein „klassisches“ Routing
- dafür: nur Pakete von Nachbarn zum Kodieren verwendet!

Wrap-up

- Network Coding
  - Vermischen von Datenströmen kann Schranken brechen!
- Data Gathering + Network Coding
  - *eine* Antwort auf die Frage, wie man Redundanz in Messdaten ausnutzt
  - bei Korrelation naher Knoten enorme Einsparung mit verhältnismäßig wenig Aufwand
  - viele Varianten denkbar!

Literatur

1. P. von Rickenbach, R. Wattenhofer. *Gathering Correlated Data in Sensor Networks*. In: ACM Joint Workshop on Foundations of Mobile Computing (DIALM-POMC), 2004